Heterogeneity and Endogenous Compliance: Implications for Scaling Class Size Interventions*

Karun Adusumilli Francesco Agostinelli Emilio Borghesan

September 2026

Abstract

This paper examines the scalability of the results from the Tennessee Student-Teacher Achievement Ratio (STAR) Project, a prominent educational experiment. We explore how the misalignment between the experimental design and the econometric model affects researchers' ability to learn about the intervention's scalability. We document heterogeneity in school-level compliance with class-size reduction that is more extensive than previously acknowledged and discuss its consequences for the evaluation of the experiment. Guided by this finding, we implement a new econometric framework incorporating heterogeneous treatment effects and endogenous class size determination. We find that the effect of class size on test scores differs considerably across schools, with only a small fraction of schools having significant benefits from reduced class sizes. We discuss the challenges this poses for the intervention's scalability and conclude by analyzing targeted class-size interventions.

Keywords: Scalability; Class Size Interventions; Treatment Effects Heterogeneity.
JEL Classification: C51, H52, I2, J13.

*We thank Michael Dinerstein for useful comments and suggestions. We also thank the participants of the workshops and seminars where we presented our work. Adusumilli: Department of Economics, University of Pennsylvania, 133 S 36th St, Philadelphia, PA 19104, USA (email: akarun@sas.upenn.edu). Agostinelli: Department of Economics, University of Pennsylvania, 133 S 36th St, Philadelphia, PA 19104, USA (email: fagostin@sas.upenn.edu). Borghesan: Department of Economics, University of Michigan, 611 Tappan Ave, Ann Arbor, MI 48109, USA (email: borghesa@umich.edu).

1 Introduction

Whether reducing class size raises academic performance is a longstanding debate in empirical economics and policy discussions (Mishel and Rothstein 2002). While early observational studies pointed towards a small or null effect of class size, results from the Tennessee Student Teacher Achievement Ratio (STAR) experiment suggested otherwise. The experiment, conducted between 1985 and 1989, randomized kindergarten students at participating public schools into one of three class types: small, regular, and regular with a teacher's aide. Multiple evaluations of STAR demonstrated significantly higher test scores for students attending the small class type, leading many researchers to conclude that class size reductions generated causal gains in student learning (Folger and Breda 1989, Finn and Achilles 1990, Word et al. 1990, Schanzenbach 2006).

The findings from STAR inspired policymakers in various states to implement large-scale class size reductions. But these policies, such as Tennessee's Project Challenge in 1989 and California's Class Size Reduction law in 1996, saw disappointing results.[1] Prior research has shown how the null effects of these two policies can be explained by either a decline in teacher quality or by insufficient class size reductions in schools (Hippel and Wagner 2018; Jepsen and Rivkin 2009). Our paper instead focuses on an overlooked aspect of the debate surrounding STAR and the scaling of experimental interventions. We ask whether and to what extent it is possible to learn about the scalability of class-size polices from the STAR experimental data itself.

We argue that the reduced-form evaluation and the two-stage least squares (2SLS) estimator commonly used in prior analyses have limited researchers' ability to learn about the effectiveness of class size policies, because these estimators are poorly-aligned with the STAR experimental design. STAR randomized students into different class types. But, class types are not the same as class sizes, and school principals had some discretion in choosing target sizes for each class type. We document substantial heterogeneity in compliance, with different schools targeting both heterogeneous sizes (doses) and heterogeneous class-size reductions between treatment and control arms. In this context, reduced-form estimates conflate heterogeneous doses and impact effects. The 2SLS estimator identifies a weighted average of school-specific effects, but these weights depend on each school's endogenous compliance with the experimental design. Neither estimator identifies the policy-relevant effect of class size on test scores, which is crucial for

[1] Project Challenge produced few benefits (Achilles et al. 1995; Hippel and Wagner 2018), and initial analysis of California's law found no effect of the statewide average reduction in class size of ten students (Bohrnstedt and Stecher 1999, 2002; Stecher and Bohrnstedt 2000).

scaling class-size initiatives.

We therefore develop and apply a new econometric method that is better-aligned with the realities of STAR's implementation. We allow class size targets and the heterogeneous effects of class size on test scores to be jointly determined by a low-dimensional set of latent parameters. We classify schools into groups based on similar values of these latent parameters. The classification works so that class size targets and impact effects are independent within each group, but they may vary across groups, thereby allowing for grouped selection on unobserved gains. Specifically, we use a Grouped Random Effects (GRE) methodology to simultaneously group schools and estimate parameters that govern the joint distribution of compliance behavior and impact effects (Adusumilli 2020). The number of groups is determined through an information criterion (BIC). BIC selects three groups as optimal.

Our method uncovers sharp differences in the effectiveness of class size reductions across schools. Nearly all of the gains from reducing class size in STAR are driven by 29% of the schools in the sample. In this set of schools, a one-student reduction in class size causes a 0.09 standard deviation (sd) improvement in test scores. The relationship between class size and test scores is modest in the remaining 71% of schools. In fact, if the 29% of highly sensitive schools had been omitted from the experiment, 2SLS would have failed to detect any causal effect of class size on test scores.

We also find evidence of heterogeneous compliance. The average school under-complied with the intended experiment by creating class-size reductions that were too small. Additionally, schools with larger impact effects in absolute magnitude under-complied still further. We simulate the counterfactual *intended* implementation of STAR – with treated class sizes ranging from 13 to 17 students, control class sizes of between 22 and 25 students, and uniform compliance across schools – and find that it would have generated 18% larger test score gains.

Our analysis of the experimental data reveals both the challenges and opportunities in learning about scaling from the STAR intervention. Potential challenges include marked heterogeneity in impact effects across schools and the fact that schools with larger impact effects, whether positive or negative, exhibited lower compliance in reducing class sizes. We decompose the policy-relevant treatment effect introduced in Heckman and Vytlacil (2001) into weighted sums of group-specific impact effects and group-specific compliance. A scaled-up version of the STAR experiment could therefore be undermined if a new population has a small proportion of schools with high impact effects (different weights) or low compliance.

We show how researchers can forecast these weights in new populations by using site-specific covariates to predict group membership as uncovered by the GRE model. Population heterogeneity can then turn into an opportunity rather than a challenge. If schools with high impact effects and compliance can be precisely identified, targeted interventions emerge as a cost-effective alternative to universal policies. To demonstrate this, we simulate a version of the STAR experiment that reduces class size at only the schools with the highest impact effects. Such an intervention would have generated test score gains that were 10.3% larger than the actual experiment while reducing the Black-white test score gap by 76.7% as much as the original experiment. Nevertheless, policymakers would struggle to identify these high impact effect schools from the STAR data. Although our approach provides tools that can identify sites with high impact effects or compliance, observable covariates are not terribly predictive in STAR.

Related Literature. An extensive empirical literature uses observational and quasi–experimental designs to estimate the effect of class size on students' educational outcomes. The results are mixed, and there is debate over whether class size truly matters. For example, Card and Krueger (1992a) find that men have a higher return to schooling in states with higher quality schools (including lower pupil/teacher ratios).[2] On the other hand, Heckman, Layne-Farrar, and Todd (1995) argue that these effects vanish once the empirical model allows for nonlinear effects of school quality on students. Hanushek (1997) argues that the relationship between school quality and student achievements becomes insignificant once researchers account for the heterogeneity in family inputs. Angrist and Lavy (1999) exploit the Maimonides' rule of classroom composition in Israeli public schools and find that class-size reductions have positive effects for fourth and fifth graders, but not for third graders. However, this finding is re-evaluated in Angrist et al. (2019) with more recent data and a larger sample, and the authors report that the effect of class-size reduction is now close to 0 for all grades. Hoxby (2000) exploits idiosyncratic variation in cohort population sizes to estimate the effect of class size on student learning and finds no effect. Rivkin, Hanushek, and Kain (2005) compare the relative importance of teachers and school quality and conclude that teacher quality matters more for children's academic achievement.

Analysis of the STAR experiment appears to rebut this conclusion. Multiple evaluations of STAR find positive short- and long-term effects on test scores and educational attainment for children who were randomly assigned to smaller classes. Because of the

[2] Card and Krueger (1992b) find that the improvement in the quality of Black schools for the cohort of Southern-born men in 1960, 1970 and 1980 explained 20% of the narrowing of the Black-white earnings gap during the same period of time.

random assignment, results from STAR have been used to reinterpret previous research on the topic and to provide strong arguments in support of policy proposals to reduce pupil/teacher ratios (see Krueger 1999, Krueger and Whitmore 2001, Schanzenbach 2006, Chetty et al. 2011, Schanzenbach 2016). We argue, however, that these prior analyses of the STAR experiment do not account for compliance heterogeneity, which limits the extent to which their conclusions are informative about class size policies in other contexts.

Our work contributes to the growing literature that studies the ability of field experiments to inform policy decisions at scale. List (2022) relates the history of attempts to scale experimental interventions and highlights many potential pitfalls. Heckman (1992) discusses how endogenous selection into field experiments impacts the external validity of the estimated impacts. Gechter et al. (2023) consider the problem of a decision maker who wants to select sites for an RCT to maximize external validity. A related literature explores the sources of treatment effect heterogeneity across sites in multi-site RCTs. Raudenbush and Bloom (2015) analyze the problem of learning about the distribution of treatment effects from multi-site trials. Kline, Rose, and Walters (2022) examine the relationship between firm-level estimates of racial bias and firm- and industry characteristics. de Chaisemartin and Deeb (2024) develop techniques to use site-level characteristics to predict intent-to-treat estimates and LATEs. In a similar manner, our GRE approach leverages the multi-site nature of STAR to show how the experimental data itself can inform both the threats and opportunities to scale class size policies in new contexts.

On the methodological side, our work is related to the large and growing literature on clustering methods and the EM algorithm. Bonhomme and Manresa (2015) and Bonhomme, Lamadon, and Manresa (2022) introduce Grouped Fixed Effects (GFE) for panel data models. GFE categorizes units into groups so that each group has the same value of unobserved heterogeneity. Grouped Random Effects instead allows the distribution of unobserved heterogeneity to vary across groups. GFE is a special case of GRE, in which the random parameters are constrained to have no variance.

We use GRE over GFE for three reasons: First, it lets us jointly model school-specific class size targets and treatment effects, which allows for correlation between compliance and impact effects across schools. Second, GRE estimates are more precise as they require fewer groups. GRE clusters observations so that the unobserved heterogeneity is approximately independent of the covariates within each group, which is weaker than requiring unobserved heterogeneity to be constant within each group as in GFE. Hence, it requires fewer groups, or, equivalently, produces smaller bias with the same number of groups. It also allows us to use posterior averaging to get better estimates of the marginal

effects. Third, GRE simultaneously computes both the group assignments and estimates of group-specific parameters. By contrast, Bonhomme, Lamadon, and Manresa (2022) suggest a two-step method, which requires specifying moments for the first-stage group assignments and can be less accurate than one-step methods. Bonhomme and Manresa (2015) suggest a one-step method for linear panel data models, but our setting is non-linear due to the need to model class sizes. A GFE approach would fail to capture the correlation between treatment effect heterogeneity and class size targets that can cause selection bias and sorting on gains.

The GRE approach uses a modified version of the EM algorithm for computation, termed EAMP. This algorithm is closely related to recent developments in computer science on variational inference (see, e.g., Blei, Kucukelbir, and McAuliffe 2017), and the interpretation of EM as a variational optimization problem (Neal and Hinton 1998).

The Grouped Random Effects framework shares similarities with empirical Bayes estimation common in the literature on teacher value-added (Chetty, Friedman, and Rockoff 2014), neighborhood fixed effects (Bergman et al. 2024), and firm-level estimates of racial discrimination (Kline and Walters 2021). In these papers, a Gaussian prior is used to shrink individual estimates in order to achieve lower mean squared error in the aggregate. Our GRE method employs Gaussian priors for the treatment effects as well, but we allow the prior parameters to vary across groups and we shrink school-specific treatment effects towards group-specific priors. This allows for greater flexibility in capturing complex patterns of unobserved heterogeneity. Additionally – and this is different from the empirical Bayes approach – we simultaneously shrink the estimates of both school-specific treatment effects and first stage compliance. Empirical Bayes estimation can thus be considered a special case of our approach.

2 Evaluation of STAR with Endogenous Class Size Reduction

2.1 The Tennessee STAR Experiment

Project STAR was a four-year longitudinal study of elementary school children in Tennessee between 1985 and 1989. In 1985, schoolchildren entering kindergarten at seventy-nine participating elementary schools were randomly allocated to one of three class types: small, regular, and regular with the addition of a teacher's aide. Teachers were randomly assigned to separate classes. The target size for the small classes was between thirteen and seventeen students, while the target size for the other two class types was between twenty-two and twenty-five students.

The experimental design called for students to remain in the same class type through the third grade. However, for various reasons—such as sample attrition, behavioral challenges, and parental complaints—many students who were initially assigned to a small class later transitioned to a different class type or withdrew from the experiment entirely. Of the 1,900 children initially assigned to a small class in kindergarten, only 857 (45%) attended a small class for all four years.[3] For this reason, we analyze the effects of class size on academic performance in kindergarten only, before any potentially endogenous reallocations of students to different class types may have taken place.[4,5]

At the end of each year of the experiment, students were given a range of cognitive and noncognitive evaluations. We focus on three cognitive tests administered to the students at the end of kindergarten: the Stanford Achievement Tests (SAT) in math, reading, and word skills. The SAT uses item response theory to facilitate comparisons across students and across years for the same student. Our dependent variable is a simple average of scores on these three exams. For the 2.3% of students with scores on only one or two exams, we construct their mean score as the average of the available scores.

Project STAR collected additional demographic information on students and their teachers. These include information on student race, gender, whether students qualified for a free or reduced price lunch (a typical proxy for low family income), absences from school, as well as teacher demographic information and qualifications. Krueger (1999) shows that the treatment and control groups in kindergarten do not differ by these observable characteristics, suggesting that the initial randomization of students into classroom types was not compromised.

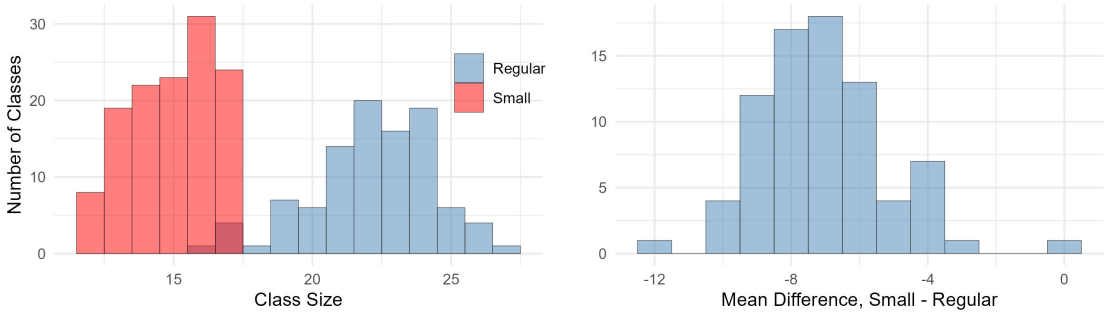
Figure 1a illustrates the distribution of class sizes in kindergarten, where the smaller "treated" classes comprised twelve to seventeen students, and the regular "control" classes ranged from sixteen to twenty-seven students. Despite the binary classification created by the experiment, class size is a polychotomous treatment variable, and we will use the

[3] Ding and Lehrer (2010) analyze the consequences of nonrandom attrition after Kindergarten for the estimation of dynamic treatment effects.

[4] Krueger (1999) provides information about attrition prior to kindergarten enrollment. He writes that 12% of students who were listed on the enrollment logs before the start of school were not enrolled in the school in the fall. The figures by class type are 10.4% of students assigned to small classes, 12.2% assigned to regular + aide classes, and 14.3% assigned to regular classes. The t-statistic for the difference between the regular and small class attrition rates is $t = 1.86$. Between 2 and 4% of students may have withdrawn from STAR schools because they were not assigned to a small class. See also Boyd-Zaharias et al. 2007 for details on the STAR experimental design.

[5] Prior research has found that assignment to the Regular + Aide group had no discernible effect on academic performance, so we omit this group from the analysis and concentrate instead on the small- and regular-sized class types (Finn and Achilles 1990, Word et al. 1990, Folger and Breda 1989). Appendix D shows that we find similar results when pooling together regular and regular + aide classrooms.

Figure 1: Distribution of Class Sizes and Class-Size Reductions



(a) Distribution of Class Sizes

(b) Compliance Heterogeneity

Panel (a) plots the number of kindergarten classes of each size for regular and small classes in the Tennessee STAR experiment. Panel (b) plots the average difference in size between small and regular classrooms at each school.

term dose throughout the text to stress this. Figure 1b illustrates the heterogeneity in dose reductions across schools. Although the average reduction in class size between control and treatment classrooms is seven students, the figure demonstrates that dose reductions varied considerably across schools, ranging between zero and twelve students. The next two sections address (i) how this heterogeneous compliance influences the interpretability of past estimates of class-size effects from the STAR experiment; and (ii) the degree to which heterogeneous compliance affects the STAR experiment's ability to inform the scalability of class-size reductions in new contexts.

2.2 The Implications of the Endogenous Class Size Reduction

The first wave of studies evaluated the reduced form of the STAR randomization by estimating specifications like the following:

$$y_{is} = \beta_0 + \beta_1 Small_{is} + \beta_2 Aide_{is} + \eta_s + \varepsilon_{is} \ , \tag{1}$$

where $Small_{is}$ is a binary variable that equals one if student i attends a small class, and $Aide_{is}$ is a binary variable that equals one if the student attends a regular-sized class with a teacher and a teacher's aide. The omitted category is a regular-sized class with a teacher but no aide. This specification compares the average test scores of students who, due to the experiment, randomly attended different class types.

Since the experiment did not specify the exact size for each type, principals were free to choose from different class sizes within a targeted range for each treatment arm. This rel-

atively free choice of class sizes plus some deviations from the targeted ranges generates the heterogeneity in dose reductions seen in Figure 1b.

Dose heterogeneity among students who attended the same class type makes generalizing the reduced form findings difficult. An analogy to randomized controlled trials in medicine serves to illustrate this point. Consider an experiment that aims to test the effectiveness of a single-use medicine by randomly assigning patients to two groups. Patients in the treatment group may select the dose they would like to be administered. Patients in the control group may also select the dose as long as it was weakly less than the minimum dose received by anyone in the treatment group. It would not then be possible to determine the effectiveness of the treatment from the reduced form of this experiment because it confounds the effectiveness of the medicine with the differences in dose intensity chosen by individuals within and between treatment and control groups. The reduced form of the STAR experiment has the same limitation: it confounds class size effectiveness with heterogeneity in dose reductions across schools.

A second wave of studies on STAR aimed to produce more generalizable results by estimating education production functions using randomization into a small class as an instrument for the class size dose within a 2SLS framework (Krueger 1999). The within-school class-type randomization, $Z_{is} \in \{0, 1\}$, serves as an instrument for class size, n_{is}, in the following model:

$$\begin{aligned} n_{is} &= \mu_s + \delta_s Z_{is} + \nu_{is} \\ y_{is} &= \eta_s + \beta_{is} n_{is} + \epsilon_{is} \ , \end{aligned} \tag{2}$$

where we omit controls for ease of interpretation. Researchers using 2SLS assume that the within-school randomization is mean independent of unobserved determinants of test scores, ϵ_{is}:

Assumption 1 (No Selection on Unobservables). *The STAR randomization between small and large class types $Z_{is} \in \{0, 1\}$ generates random student compositions between different class types: $E[\epsilon_{is}|Z_{is}, s] = E[\epsilon_{is}|s]$.*

It is well known that the causal interpretation of 2SLS estimates with ordered multi-valued treatment doses is complicated by heterogeneous dose reductions across units, as discussed in Angrist and Imbens 1995, Rose and Shem-Tov 2021, and in the context of essential heterogeneity in Heckman, Urzua, and Vytlacil (2006). In the context of STAR, 2SLS fails to identify the policy-relevant marginal effect of class size on test scores even

under Assumption 1. Proposition 1 shows that it instead identifies a weighted average of school-specific effects, where the weights depend on schools' endogenous dose reductions.

Proposition 1. *Suppose Assumption 1 holds, and that school-specific treatment effects are linear. Moreover, suppose that the original STAR randomization is relevant, i.e., it generates significant differences in class size within each school. The 2SLS estimator of model (2) identifies a weighted average of marginal effects, where the weights depend on the endogenous compliance behavior of schools with respect to the class size reduction in the experimental setting:*

$$\beta_1^{2sls} = \sum_{s \in \mathcal{S}} \beta_s \frac{\phi_s P_{zs}(1 - P_{zs}) \left(\mathbb{E}[n_{is}|s, Z_{i,s} = 1] - \mathbb{E}[n_{is}|s, Z_{i,s} = 0]\right)}{\sum_{s \in \mathcal{S}} \phi_s P_{zs}(1 - P_{zs}) \left(\mathbb{E}[n_{is}|s, Z_{i,s} = 1] - \mathbb{E}[n_{is}|s, Z_{i,s} = 0]\right)} . \tag{3}$$

Proof. See Appendix B. ■

In the above equation $\beta_s = \mathbb{E}[\beta_{is}|s]$, ϕ_s is the population fraction of students attending school s, $P_{zs} = \mathbb{P}(Z_{is} = 1 \mid s)$, and $\mathbb{E}[n_{is}|s, Z_{i,s} = 1] - \mathbb{E}[n_{is}|s, Z_{i,s} = 0]$ represents the expected difference in treated and control class sizes at school s.

The average marginal effect of class size on test scores, $\mathbb{E}[\beta_{is}]$, is the main policy-relevant parameter of interest, but Proposition 1 shows that 2SLS will only identify $\mathbb{E}[\beta_{is}]$ in two special cases: (i) when impact effects are homogeneous across schools, namely $\beta_s = \beta$ for all $s \in \mathcal{S}$; or (ii), when schools have equal expected levels of compliance regarding dose reductions: $\mathbb{E}[n_{is}|s, Z_{i,s}] = \mathbb{E}[n_{is}|Z_{i,s}]$ and $\mathbb{P}(Z_{is} = 1 \mid s) = \mathbb{P}(Z_{is} = 1)$. Both cases impose strong restrictions on either impact heterogeneity or on schools' objective functions. In the first case, an extensive literature has demonstrated the existence of heterogeneous treatment effects in educational settings (Carneiro, Heckman, and Vytlacil 2011, Walters 2018, Borghesan and Vasey 2024). The second case will be violated if, for example, schools choose a particular reduction in class size between small and regular class types, thereby self-selecting into a particular intensity of treatment on the basis of their unobserved gains from the experiment.

Beyond these special cases, the 2SLS estimand reflects a weighted average of school-specific marginal effects of class size on test scores. The weights, as defined in (3), are influenced by three key factors, each of which may be correlated with the school-specific effect of class size on test scores. They are largest in schools with high enrollment (high ϕ_s), in schools that have an equal proportion of treated and control students, which maximizes $P_{zs}(1 - P_{zs})$, and in schools that generate a significant difference in class size

Table 1: Example of How IV Weights Are Determined

	N	Treated	Control	ϕ_s	$P_{zs}(1-P_{zs})$	Dose Reduction	Weight
A	45	15	30	0.3	0.22	-15	0.35
B	45	20	25	0.3	0.25	-5	0.13
C	60	15, 15	30	0.4	0.25	-15	0.52

The table provides an example of the control and treatment class sizes created by three hypothetical schools. N is the enrollment at the school, the second and third columns refer to the treated and control classroom sizes created, and the remaining columns correspond to elements of the IV weights in (3), with $P_{zs} := \mathbb{P}(Z_{is} = 1 \mid s)$.

between treatment and control classrooms.[6] The last two factors are endogenous in the STAR experiment.

Table 1 provides an example with three schools to demonstrate how endogenous class size choices may affect the weights in the 2SLS estimand. School A has 45 students and creates a treatment class of 15 and a control class of 30. School B has 45 students and creates a treatment class of 20 and a control class of 25. School C has 60 students and creates two treatment classes of 15 students and one control class of 30 students. School A has a low population, a low treated fraction of $1/3$, but a large dose reduction between treated and control classrooms. School B has a higher fraction of treated students than school A but has a lower dose reduction. Finally, School C is associated with a high value for all components that affect the 2SLS weights. School C is then overweighted in the 2SLS estimand relative to its size, while school B is significantly underweighted.

2.3 Lessons from STAR for Class-Size Reduction Policies at Scale

Is the compliance heterogeneity of STAR a flaw or a virtue? Although endogenous compliance poses a challenge for conventional estimators, it can also generate insights into the factors that influence the adoption of policies at scale (DellaVigna, Kim, and Linos 2024). In STAR, each school represents a distinct experiment, which allows the researcher to study how impact heterogeneity and dose reductions vary systematically across schools. This approach aligns with the recommendation of Al-Ubaydli, List, and Suskind (2020) that researchers leverage multi-site trials to investigate the variability in

[6]The weights will be nonnegative at any school where the average treated classroom has weakly fewer students than the average control classroom. This was the case with all schools in the STAR experiment.

program impacts across diverse populations and situational contexts. They argue that "the design of multi-site trials can provide empirical content into why effects might not scale and give empirical hints about where more research is necessary before scaling."

One way to learn about impact heterogeneity in the STAR experiment would be to estimate the relationship between class size and test scores separately by school. Even if impact heterogeneity and dose reductions were correlated *across schools*, the within-school random assignment of students and teachers to classes of varying sizes would produce unbiased estimates of school-specific impact effects. However, limited variation in class size within each school would then lead to noisy estimates, amplifying the scalability threat posed by low-powered studies (Al-Ubaydli, List, and Suskind 2020).[7] Such a strategy would be suboptimal for researchers and policymakers who intend to extrapolate from the STAR experiment to class-size policies in other contexts. We therefore propose a new econometric framework that aims to balance the twin attractions of statistical efficiency and allowing for unobserved heterogeneity in impact effects and compliance. Our approach will thus shed insight on how these two forces affect the scalability of the STAR experiment.

3 From STAR Randomization to Policy-Relevant Effects

In this section, we introduce a model aimed at uncovering heterogeneous effects of class size on test scores in the presence of potentially endogenous compliance. Our method clusters schools into a fixed number of groups, $k = 1, \ldots, K$, that share a common set of underlying parameters which collectively govern both the distribution of class sizes within each school and the marginal effects of class size on test scores.

3.1 Test Score Model

Test scores are generated by the following equation for student i in school s belonging to group k:

$$\begin{aligned} y_{isk} &= \eta_s + \beta_{isk} n_{isk} + x_{isk}'\theta + \epsilon_{isk} \ , \\ \beta_{isk} &\sim N(\mu_k, \Sigma_k) \ , \\ \epsilon_{isk} &\sim N(0, \sigma^2_{\epsilon,k}), \end{aligned} \tag{4}$$

[7] The median school in the experiment features two distinct classes: one small and one regular.

where n_{isk} is individual i's class size and x_{isk} is a vector of observable student characteristics.[8] In the model, impact effects are heterogeneous and are assumed to follow a normal distribution with mean, μ_k, and variance, Σ_k, that are allowed to differ by group. The term μ_k captures the expected effect of a 1-unit increase in class size on the outcome for a randomly selected student attending a school in group k. Student demographics, x_{isk}, instead have non-random effects on test scores. Test scores additionally depend on a full set of school fixed effects, η_s, and a student-specific shock, ϵ_{isk}, with mean 0 and a variance $\sigma^2_{\epsilon,k}$ that is allowed to vary by group. The model therefore incorporates group-specific heteroskedasticity.

The random assignment of teachers and students to classrooms is implicitly imposed via two modeling choices. First, the distribution of treatment effects is school- but not classroom-specific, so that what matters for the marginal effect of class size on test scores is the match between the student and the school. Second, assignment to the treatment or control group only affects test scores through its impact on classroom size, n_{isk}.

3.2 Class Size Model

Each school is associated with a vector of class sizes, $\boldsymbol{n}_{sk}^{(t)}$ and $\boldsymbol{n}_{sk}^{(c)}$, for the treated and control groups.[9] We posit that these class sizes are generated according to the following multinomial model:

$$\begin{aligned} \boldsymbol{n}_{sk}^{(c)} &\sim \text{multinomial}(p_{sk}^{(c)}) \text{ for all } k \text{ in the control group,} \\ \boldsymbol{n}_{sk}^{(t)} &\sim \text{multinomial}(p_{sk}^{(t)}) \text{ for all } k \text{ in the treatment group.} \end{aligned} \tag{5}$$

Here, $p_{sk}^{(c)} \equiv \{p_{sk1}^{(c)}, \ldots, p_{skL}^{(c)}\}$, $p_{sk}^{(t)} \equiv \{p_{sk1}^{(t)}, \ldots, p_{skM}^{(t)}\}$ are school-specific probability distributions over the size of each class. The support for these distributions is the same as the support for the class size vectors, $\boldsymbol{n}_{sk}^{(t)}$ and $\boldsymbol{n}_{sk}^{(c)}$, and is equal to their observed support in the data: $\{12, \ldots, 17\}$ and $\{16, \ldots, 27\}$ respectively (so that $L = 6$ and $M = 12$).

The model assumes that class sizes are drawn from school-specific multinomial distributions, with multiple classes in the same school of a particular type (either treatment

[8] We depart from Krueger (1999) in using raw scores rather than percentile ranks as the outcome variable. Causal models of the ranks of a dependent variable do not admit a traditional interpretation of coefficients as marginal effects. In addition, our model assumes that outcomes are independently distributed conditional on group membership, while the use of ranks would induce dependence across groups. Finally, a model of test score ranks is inappropriate for many forms of counterfactual analysis, as a reduction in class size for every student may have beneficial effects on academic outcomes without causing any discernible effect on ranks.

[9] $\boldsymbol{n}_{sk}^{(t)}$ and $\boldsymbol{n}_{sk}^{(c)}$ represent vectors of counts of the number of classes in school s of each size. The class size experienced by each student in equation (4), n_{isk}, is a scalar integer-valued variable.

or control) representing independent draws from the same distribution. One can intuitively think of $p_{sk}^{(c)}$ and $p_{sk}^{(t)}$ as denoting the target proportions of control and treatment classroom sizes for each school. The observed class sizes are then random deviations from this target. Modeling $p_{sk}^{(c)}$ and $p_{sk}^{(t)}$ as school-specific rather than classroom-specific is consistent with the random assignment of students and teachers to classrooms. We do not model the assignment of treatment and control status to the classroom, since these were determined randomly in the experiment.

Importantly, we model class size levels and not simply the difference in class size between treatment and control classrooms. It is quite plausible that schools with different control class sizes are associated with different impact effects, causing selection bias. Sorting on gains could result if schools with larger impacts choose a larger (or smaller) class-size reduction. By modeling the class size levels for both treatment and control class types, we are able to account for both selection bias and sorting on gains.

Note that we do not take the number of students in each school as given. It is endogenous in our model, determined by the sizes of each class in the school. Endogenizing school size in this manner is useful if school-level factors determining cohort size are correlated with the impact effects, β_{isk}.

3.3 Endogeneity and Grouped Random Effects

We adopt a Grouped Random Effects approach, following Adusumilli (2020), to simultaneously estimate (4) and (5) and to group schools that have a common set of underlying parameters that jointly determine class size and impact effects. These common parameters, ρ_k, are constant within each group, and they serve as priors over each group's random parameters, $(\beta_{isk}, p_{sk}^{(c)}, p_{sk}^{(t)})$. Let $k \in 1, \ldots, K$ denote the set of groups, and $w_s(k)$ the group assignment, with $w_s(k) = 1$ if school s is in group k. For the prior on the treatment effect coefficients, $\boldsymbol{\beta} \equiv \{\boldsymbol{\beta}_{isk} : s = 1, \ldots, S; s(i) = s; w_s(k) = 1, k = 1, \ldots, K\}$, we specify

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) := \prod_{s} \prod_{i:s(i)=s} \prod_{k} N(\beta_{isk}|\mu_k, \Sigma_k)^{w_s(k)}, \tag{6}$$

so that the student-specific marginal effects, β_{isk}, are random draws from a group-specific normal distribution with mean μ_k and variance Σ_k. This assumes that the distribution of unobserved student and school characteristics affecting the relationship between class size and test scores is the same across all schools within the same group.

The school-specific multinomial probabilities for class size, $\boldsymbol{p}^{(c)} \equiv \{p_{sk}^{(c)}\}_{s=1}^{S}, \boldsymbol{p}^{(t)} \equiv \{p_{sk}^{(t)}\}_{s=1}^{S}$,

are assumed to have the following group-specific Dirichlet priors:

$$\pi(\boldsymbol{p}^{(c)}|\boldsymbol{\eta}^{(c)}) = \prod_s \prod_k \text{Dirichlet}(p_{sk}^{(c)}|\eta_k^{(c)})^{w_s(k)},$$
$$\pi(\boldsymbol{p}^{(t)}|\boldsymbol{\eta}^{(t)}) = \prod_s \prod_k \text{Dirichlet}(p_{sk}^{(t)}|\eta_k^{(t)})^{w_s(k)}, \tag{7}$$

where $\boldsymbol{\eta}^{(c)} \equiv \{\eta_k^{(c)}\}_{k=1}^K, \boldsymbol{\eta}^{(t)} \equiv \{\eta_k^{(t)}\}_{k=1}^K$ are group-specific, and Dirichlet$(p|\eta)$ denotes the pdf of the Dirichlet distribution with parameter η evaluated at p. The Dirichlet distribution is chosen both for computational tractability, as it is the conjugate prior for the multinomial distribution, and because it is a distribution over probability distributions on the unit simplex. This makes it an appropriate choice to model probability vectors that govern the sizes of control and treatment classes at each school. Overall, the set of parameters to be estimated are $\rho_k := \{\mu_k, \Sigma_k, \eta_k^{(c)}, \eta_k^{(t)}\}$ for $k = 1, \ldots, K$ and θ.

Implicit in our probability model is the following assumption about the distribution of random parameters:

Assumption 2 (Within-Group IID Random Effects). *We assume that the random coefficients of class size on test scores, β_{isk}, follow identical group-specific distributions that are independent of class size: $\beta_{isk} \perp\!\!\!\perp n_{sk}^{(t)}, n_{sk}^{(c)} \mid k$ for all i in school s.*

Assumption 2 means that GRE allows for correlation between class size and test scores only across groups. This is driven by correlation between μ_k, which governs the marginal effect of class size on test scores, and $\eta_k^{(t)}$ and $\eta_k^{(c)}$, which govern the distributions of treated and control class sizes within each group. However, conditional on the group, the realized class size for each student is independent of its marginal effect on test scores.

The within-school random assignment of students and teachers to classes ensures that assumption 2 would be satisfied if the number of groups were set equal to the number of schools. As we cluster schools together to gain efficiency, a crucial question is whether this comes at the expense of introducing bias. Bias would occur if class size effects were correlated with the class size reduction even within groups. We determine the number of groups using a model selection criterion that balances improvements in the likelihood with the cost of additional parameters. Intuitively, if increasing the number of groups causes negligible changes in the maximized likelihood, then any within-group correlation between treatment effects and compliance must have been weak at the smaller group size.

3.4 Probability model

Let $\boldsymbol{y} \equiv \{y_{isk}\}_{i,s,k}$ and $\boldsymbol{x} \equiv \{x_{isk}\}_{i,s,k}$ denote the vectors of test scores and covariates, and group the variances of the test score disturbances as $\boldsymbol{\sigma}_\epsilon^2 \equiv \{\sigma_{\epsilon,k}^2\}_k$. Given the treatment effect heterogeneity $\boldsymbol{\beta}$, and the class size vectors $\boldsymbol{n}^{(c)} := \{n_{isk}^{(c)}\}_{i,s,k}$, and $\boldsymbol{n}^{(t)} := \{n_{isk}^{(t)}\}_{i,s,k}$, equation (4) implies that the log-likelihood of the test score model is

$$\begin{aligned}\ln p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{n}^{(c)}, \boldsymbol{n}^{(t)}, \boldsymbol{x}, \boldsymbol{\sigma}_\epsilon^2, \theta) &:= \sum_s \sum_{i:s(i)=s} \ln p(y_{isk}|\beta_{isk}, n_{isk}, x_{isk}, \sigma_{\epsilon,k}^2, \theta, s) \ , \\ &:= \sum_s \sum_{i:s(i)=s} -\frac{1}{2}\ln \sigma_{\epsilon,k(s)}^2 - \frac{(y_{isk} - \beta_{isk} n_{isk} - x_{isk}'\theta - \eta_s)^2}{2\sigma_{\epsilon,k(s)}^2} \ .\end{aligned}$$

Similarly, the multinomial class size model in (5) implies that the distribution of the sizes of treatment and control classrooms is given by

$$\ln p(\boldsymbol{n}^{(c)}|\boldsymbol{p}^{(c)}) := \sum_{s=1}^{S} \sum_{g=1}^{N^{(c)}(s)} \sum_{k=1}^{K} w_s(k) \left\{ \sum_{l=0}^{L-1} \mathbb{I}\{n_{skg}^{(c)} \equiv 12 + l\} p_{skl}^{(c)} \right\} \ , \tag{8}$$

$$\ln p(\boldsymbol{n}^{(t)}|\boldsymbol{p}^{(t)}) := \sum_{s=1}^{S} \sum_{g=1}^{N^{(t)}(s)} \sum_{k=1}^{K} w_s(k) \left\{ \sum_{m=0}^{M-1} \mathbb{I}\{n_{skg}^{(t)} \equiv 16 + m\} p_{skm}^{(t)} \right\} \ , \tag{9}$$

where $N^{(c)}(s)$ and $N^{(t)}(s)$ are the observed number of control and treatment classes in school s. The expressions in (8) and (9) show that the log-likelihood of obtaining the vector of control and treated class sizes, $\boldsymbol{n}^{(c)}$ and $\boldsymbol{n}^{(t)}$, is equal to the sum, over all groups, of an indicator for group membership multiplied by the school-group-specific multinomial probabilities for the observed class sizes.

Conditional on the unobserved heterogeneity, $(\boldsymbol{\beta}, \boldsymbol{p}^{(c)}, \boldsymbol{p}^{(t)})$, the joint likelihood of the observations $(\boldsymbol{y}, \boldsymbol{n}^{(c)}, \boldsymbol{n}^{(t)})$ is given by

$$p(\boldsymbol{y}, \boldsymbol{n}^{(c)}, \boldsymbol{n}^{(t)}|\boldsymbol{\beta}, \boldsymbol{p}^{(c)}, \boldsymbol{p}^{(t)}, \boldsymbol{x}, \boldsymbol{\sigma}_\epsilon^2, \theta) := p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{n}^{(c)}, \boldsymbol{n}^{(t)}, \boldsymbol{x}, \boldsymbol{\sigma}_\epsilon^2, \theta) \cdot p(\boldsymbol{n}^{(c)}|\boldsymbol{p}^{(c)}) \cdot p(\boldsymbol{n}^{(t)}|\boldsymbol{p}^{(t)}).$$

Additionally, in view of (6) and (7), the prior distribution of unobserved heterogeneity is

$$\pi\left(\boldsymbol{\beta}, \boldsymbol{p}^{(c)}, \boldsymbol{p}^{(t)}|\boldsymbol{\gamma}, \boldsymbol{\eta}^{(c)}, \boldsymbol{\eta}^{(t)}\right) := \pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) \cdot \pi(\boldsymbol{p}^{(c)}|\boldsymbol{\eta}^{(c)}) \cdot \pi(\boldsymbol{p}^{(t)}|\boldsymbol{\eta}^{(t)}),$$

where $\boldsymbol{\gamma} = \{\mu_k, \Sigma_k\}_k$ denotes the collection of group-specific parameters determining the effects of class size on student test scores.

4 Estimation

Let $\boldsymbol{\alpha} := (\boldsymbol{\beta}, \boldsymbol{p}^{(c)}, \boldsymbol{p}^{(t)})$, $\boldsymbol{\beta}_s := \{\beta_{isk} : s(i) = s\}$, $\boldsymbol{\alpha}_s := (\boldsymbol{\beta}_s, p_s^{(c)}, p_s^{(t)})$, $\boldsymbol{\rho} := (\boldsymbol{\gamma}, \boldsymbol{\eta}^{(c)}, \boldsymbol{\eta}^{(t)})$ and $\boldsymbol{\rho}_k := (\boldsymbol{\gamma}_k, \boldsymbol{\eta}_k^{(c)}, \boldsymbol{\eta}_k^{(t)})$. Also, let $(\boldsymbol{y}_s, \boldsymbol{n}_s^{(c)}, \boldsymbol{n}_s^{(t)})$ denote the set of test scores and class sizes pertaining to school s. The GRE problem is to maximize the likelihood of the data jointly over both the group assignments and the common and group-specific parameters:

$$\begin{aligned}
&\max_{\{w_s(k)\},\theta,\{\boldsymbol{\rho}_k\}} \ln \int p(\boldsymbol{y}, \boldsymbol{n}^{(c)}, \boldsymbol{n}^{(t)} | \boldsymbol{\alpha}, \boldsymbol{x}, \boldsymbol{\sigma}_\epsilon^2, \theta) \pi\left(\boldsymbol{\alpha} | \boldsymbol{\rho}\right) d\boldsymbol{\alpha} \\
&= \max_{\{w_s(k)\},\theta,\{\boldsymbol{\rho}_k\}} \sum_{s=1}^{S} \sum_{k=1}^{K} w_s(k) \ln \int p(\boldsymbol{y}_s, \boldsymbol{n}_s^{(c)}, \boldsymbol{n}_s^{(t)} | \boldsymbol{\alpha}_s, \boldsymbol{x}, \sigma_{\epsilon,k}^2, \theta) \pi\left(\boldsymbol{\alpha}_s | \boldsymbol{\rho}_k\right) d\boldsymbol{\alpha}_s. \qquad (10)
\end{aligned}$$

Following Adusumilli (2020), we solve the above maximization problem using the Expectation, Assignment, Maximization, and Propagation (EAMP) algorithm. To use the algorithm, we first use the Donsker-Varadhan variational formula to rewrite the likelihood in (10) as follows:

$$\begin{aligned}
&\max_{\{w_s(k)\},\theta,\{\boldsymbol{\rho}_k\}} \sum_{s=1}^{S} \sum_{k=1}^{K} w_s(k) \ln \int p(\boldsymbol{y}_s, \boldsymbol{n}_s^{(c)}, \boldsymbol{n}_s^{(t)} | \boldsymbol{\alpha}_s, \boldsymbol{x}, \sigma_{\epsilon,k}^2 \theta) \pi\left(\boldsymbol{\alpha}_s | \boldsymbol{\rho}_k\right) d\boldsymbol{\alpha}_s \\
&= \max_{\substack{\{q_{sk}(\cdot)\}, \\ \{w_s(k)\},\theta,\{\boldsymbol{\rho}_k\}}} \sum_{s=1}^{S} \sum_{k=1}^{K} w_s(k) \Big\{ E_{q_{sk}(\cdot)} \left[\ln p(\boldsymbol{y}_s, \boldsymbol{n}_s^{(c)}, \boldsymbol{n}_s^{(t)} | \boldsymbol{\alpha}_s, \boldsymbol{x}, \boldsymbol{\sigma}_{\epsilon,k}^2, \theta) \right] - \\
&\qquad\qquad \mathrm{KL}\left(q_{sk}(\boldsymbol{\alpha}_s) \,||\, \pi(\boldsymbol{\alpha}_s | \boldsymbol{\rho}_k)\right) \Big\}. \qquad (11)
\end{aligned}$$

In the above equation, $q_{sk}(\cdot)$ denotes a group-specific distribution over $\boldsymbol{\alpha}_s$ and the maximization is carried out over the space of all possible distributions $q_{sk}(\cdot)$. The EAMP Algorithm proceeds by repeatedly maximizing over each of $\{q_{sk}(\cdot)\}$, $\{w_s(k)\}$, θ, and $\{\boldsymbol{\rho}_k\}$ holding other quantities fixed. This results in a sequence of four steps – Expectation, Assignment, Maximization, and Propagation – that are repeated in an iterative process until the algorithm converges. Note that we first demean all variables by their school-specific means to eliminate the school fixed effects before running the estimation algorithm. A detailed description of the algorithm is provided in Appendix A.

The EAMP algorithm generalizes the standard EM algorithm to allow for grouping of data. In addition, it splits the usual M step into a Maximization and Propagation step; the latter enables fast updates for the prior. Finally, the specific order of the steps in EAMP is chosen for computational efficiency. Since EAMP is a type of coordinate ascent algorithm, it is guaranteed to converge. However, as with the EM algorithm, it may con-

verge to a saddle point or local maximum instead of a global maximum. Consequently, to ensure convergence to a global optimum, we initialize the algorithm with multiple random starting points.

5 Empirical Findings

In this section we present the estimates of the model with controls for child gender, race (an indicator for being neither white nor Asian), and free lunch status. The EAMP algorithm takes the number of groups as given. In choosing the group size, we aim to balance the competing attractions of parsimony and allowing for richer patterns of heterogeneity. The EAMP framework allows for the number of groups to be as many as the number of schools. However, allowing for too many groups may cause the model to capture more noise than signal and render the estimates difficult to interpret. In practice, we use the Bayesian Information Criterion (BIC) to select the number of groups. BIC selects three groups as optimal.[10]

Altogether, our analysis consists of kindergarten students in the seventy-nine participating schools in the Tennessee STAR experiment who were randomly assigned to either a small or a regular class, who do not lack information on race, gender, or eligibility for free/reduced price lunch, and who have at least one exam score at the end of kindergarten. These restrictions result in an estimation sample of 3813 students. Summary statistics for the estimation sample are presented in Appendix Table C-1. The exams, which were originally measured on a scale of 0 to 1000, have been converted to a 0-100 scale. The table shows that the mean exam score is 45.30 points, and the standard deviation is 3.59 points. The minimum and maximum test scores in the sample are 28.80 and 61.53, respectively. The average student sits in a classroom with nearly 19 students including herself. 49% of students are female, 47% of students are eligible for free lunch, and 32% are neither white nor Asian, meaning that they are either Black, Hispanic, or Native American. Slightly fewer than half of students, 46.5%, are in small (treated) classrooms.

Table 2 presents the estimated parameters of the test score model, the mean (μ_k) and variance (Σ_k) of the marginal effects of class size on test scores and the variance of the unobserved test score shocks, $\sigma^2_{\epsilon,k}$. Our results reveal striking differences in the marginal effects of class size on test scores across groups. Group two, comprising twenty-three schools, exhibits substantial gains from reducing class size. A one-student reduction in class size at schools in group two causes in a 0.339 point test score increase on average,

[10] As a robustness exercise, Appendix E presents estimation results with four groups.

Table 2: Class Size Marginal Effects by Group

Group	μ_k	Σ_k	$\sigma^2_{\epsilon,k}$	Schools	Students
1	-0.068	0.042	4.933	31	1425
	(0.061)	(0.065)	(1.55)		
2	-0.339	0.235	6.442	23	1137
	(0.062)	(0.15)	(1.86)		
3	0.106	0.001	11.852	25	1251
	(0.059)	(0.07)	(2.07)		
Average Effect	-0.092	0.118	7.653		
	(0.033)	(0.048)	(0.68)		
N	3813				

The table shows the estimated model parameters governing the effect of class size on test scores. The model controls for gender, race, and free lunch status, and their coefficients are constrained to be equal across groups. The class size effects represent the effect of a one-unit increase in class size on a simple mean of three test scores. Bootstrapped standard errors from 95 bootstrapped data sets are in parentheses. Bootstrapping involves first sampling schools from the set of 79 schools that participated in the experiment and then sampling individuals within each school.

equivalent to an improvement of 0.094 standard deviations. In contrast, group three, encompassing twenty-five schools, demonstrates an insignificant but positive effect.[11] Class size has a negative, but statistically insignificant, effect at schools in group one. When averaged across all groups and students, the marginal effect of a one-student reduction in class size translates to an improvement of 0.092 points, equivalent to 0.026 standard deviations.[12]

These heterogeneous effects of class-size reductions are accompanied by heterogeneous compliance across schools when determining class sizes. Table 3 displays the estimated Dirichlet prior means for the class size model by group. In our context, they can be interpreted as a prior on the fraction of classes with each size. A higher value, all else

[11] A positive effect of class size on test scores is surprising, but the model takes no stance on the mechanism relating class size to test scores across groups. Teachers in the large classes in group 3 may have responded to the size of their classes by conducting different classroom activities than the large-class teachers in other groups. As Kindergarten incorporates play and other activities that develop social skills and emotional regulation, the children in the large classes in group 3 may have benefited from the increased interaction.

[12] Appendix Table E-2 shows that these findings are robust to a model that allows the coefficients on X to vary by group.

Table 3: Dirichlet Class Size Parameters by Group

	Treatment Class Size					
Group	12	13	14	15	16	17
1	0.06	0.15	0.09	0.22	0.25	0.22
2	0.03	0.14	0.29	0.13	0.21	0.20
3	0.08	0.17	0.16	0.17	0.26	0.15

	Control Class Sizes											
Group	16	17	18	19	20	21	22	23	24	25	26	27
1	0	0.03	0.03	0.11	0	0	0.14	0.17	0.32	0.11	0.09	0
2	0	0	0	0.03	0.12	0.33	0.32	0.07	0.10	0.03	0	0
3	0.03	0.09	0	0.06	0.06	0.12	0.15	0.25	0.15	0.03	0.03	0.03

The top panel of the table displays the Dirichlet prior means for each treatment class size by group. The bottom panel reports the Dirichlet prior means for each control class size by group. All parameters have been normalized by the sum of the estimated parameters for each group. Given this normalization, each number in the table can be interpreted as a group-specific prior on the fraction of observations in each cell. Zeros indicate that a particular group does not generate classes of that size.

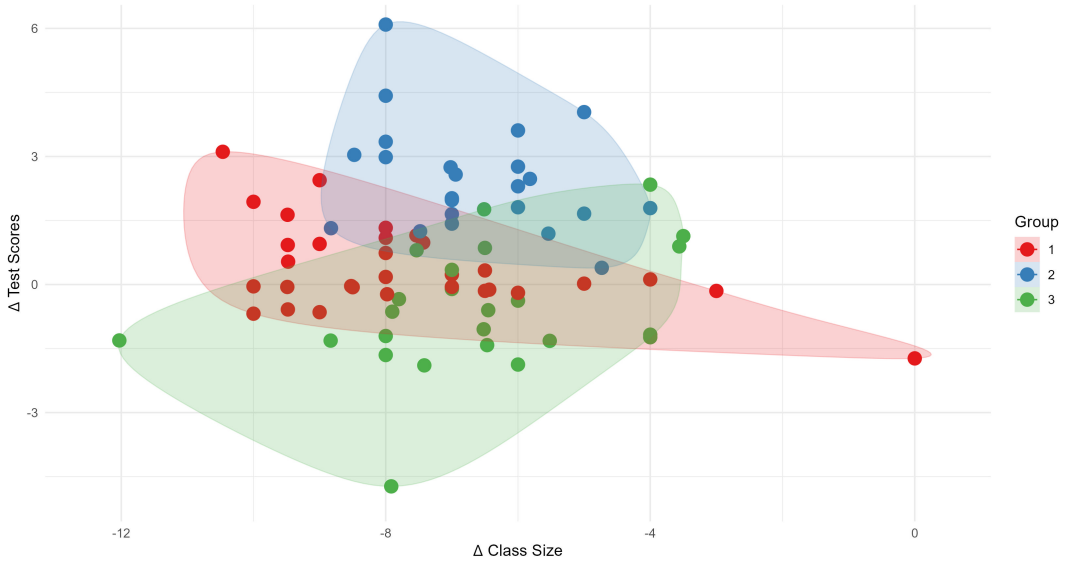
equal, indicates that schools within that group are more likely to create a classroom of that size. The table shows that schools in group 1 choose relatively large control class sizes, while group 2 has relatively small control classes, and group 3 has smaller treated class sizes. Variation in targeted treatment and control class sizes means that compliance with the STAR protocol differed across schools, with schools in group 1 creating the largest expected difference between treated and control classrooms.[13]

The EAMP algorithm thus uncovers distinct patterns of learning gains and class size reductions in the data. Figure 2 plots the within-school treatment-control difference in test scores against the within-school treatment-control difference in class sizes. The areas inhabited by Groups 2 and 3 scarcely overlap. Group 2 consists of schools where students attending small classes saw large benefits, while students in group 3 schools benefited little from smaller classes. Group 1, which has a moderate relationship between class size and test scores, is relevant for policies that reduce class size marginally from the status quo, because it is the only group with any support on small reductions in class size.

In Appendix Table C-2, we show how students in different groups differ by observable

[13]Table C-7 compares the means and standard deviations of the model's endogenous variables to the data overall and by group. The model overestimates the standard deviation of test scores, but otherwise the fit is good.

Figure 2: Test Scores and Class Size Reductions



The figure plots the within-school average difference in test scores between treated and control classrooms along the Y-axis against the average within-school difference in class size between treated and control classrooms along the X-axis. Each dot represents a school. The groups are determined by the EAMP algorithm described in section 3.

characteristics. We repeat the same analysis for teacher-specific variables in Table C-3 and for school-specific variables in Table C-4. Observable differences across groups are small, although a few are of particular interest. Group 2, which has the greatest marginal effect of class size, has the highest fraction of nonwhite students. Students in group 2 also attend school for between four and five fewer days of the school year relative to students in the other groups. Group 3, which is the only group with a positive (but insignificant) relationship between class size and student performance, has the highest fraction of white and Asian students. Table C-3 shows that schools in Groups 2 and 3, which have opposite responses to class size reductions, display similar populations of teachers. These teachers are less likely to have a masters degree but have more years of experience than group 1 teachers.

Table C-4 shows little evidence of geographic differences across the three groups. Regarding enrollment, group 2 schools tend to have the largest cohorts, while schools in group 1, which complied the most with the policy by creating large class size reductions, had the smallest average cohort size and the least race-segregated schools.[14] These find-

[14] A school is deemed to be race-segregated if over 80% of its student body belongs to a single race. According to this metric, 88% of schools in the STAR experiment were segregated.

Table 5: Heterogeneity in Marginal Effects of Class Size on Test Scores by Race and Income

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	By Race				By Income			
	Nonwhite	White & Asian	Nonwhite	White & Asian	Free lunch	No free lunch	Free lunch	No free lunch
Class size	−0.134	−0.085			−0.110	−0.093		
	(0.025)	(0.018)			(0.020)	(0.021)		
Class size × Group 1			−0.169	−0.023			−0.116	−0.032
			(0.035)	(0.026)			(0.028)	(0.029)
Class size × Group 2			−0.349	−0.346			−0.311	−0.357
			(0.050)	(0.032)			(0.038)	(0.036)
Class size × Group 3			0.114	0.115			0.098	0.123
			(0.038)	(0.033)			(0.031)	(0.038)
Constant	−0.282	0.132	−0.271	0.116	−0.705	0.632	−0.564	0.477
	(0.086)	(0.063)	(0.142)	(0.099)	(0.069)	(0.072)	(0.112)	(0.113)
Observations	1,216	2,597	1,216	2,597	1,803	2,010	1,803	2,010
R^2	0.014	0.008	0.063	0.048	0.011	0.009	0.051	0.053

Columns (1)-(2) and (5)-(6) show 2SLS regression results for four subsamples of interest: nonwhite, white & Asian, students who do not qualify for free or reduced price lunch, and those who do qualify. The excluded instrument is a binary indicator for attending a small classroom. Columns (3)-(4) and (7)-(8) show the results from OLS regressions of test scores on class size interacted with group membership, as uncovered by the EAMP algorithm, for the same subsamples. The dependent variable is a simple mean of scores on three cognitive tests. All models include school fixed effects.

ings suggest that the schools in Tennessee that deliver the greatest marginal returns to class size have more nonwhite students and larger cohorts. The results also suggest that compliance was lowest at the most segregated schools.

5.1 Individual Heterogeneity in the Impact of Class Size

Krueger (1999) found that that Black students and students qualifying for a free or reduced price lunch benefited more from smaller class sizes than did their white and wealthier peers. In Table 5, we show that these differential effects are largely driven by differences across schools rather than being a within-school phenomenon. Columns (1) and (2) display separate 2SLS regressions of test scores on class size for nonwhite and white and Asian students, while columns (3) and (4) run OLS regressions after interacting class size with group membership as recovered by the EAMP algorithm. Under identifying assumption 2, the random coefficient on class size, β_{isk}, is independent of class size conditional on group membership. Interacting class size with group membership is therefore sufficient to alleviate the endogeneity problem.

The table shows that differences in the returns to class size by race identified by the 2SLS regressions are entirely driven by differences in group one. White and Asian students

in group 1 schools do not benefit from reduced class sizes, while nonwhite students experience significant gains. There is no racial difference in the return to class size at schools in Groups 2 and 3, suggesting that individual heterogeneity in the response to class size interventions is driven at least in part by the type of school attended. Columns (5)-(8) of Table 5 repeat the same exercise for students qualifying for free lunch and those who do not. Columns (7) and (8) show that students in group 1 schools who qualify for free lunch benefit from smaller class sizes, while students from richer households do not, although these differences are not statistically significant.[15] This information is relevant for policymakers, because if their goal is to reduce the Black-white test score gap, they will want any proposed policy to include a sizeable number of group 1 schools.

5.2 Are the Effects of Class Size Nonlinear?

An important assumption of our model is that class size has a linear effect on test scores in each group.[16] We examine this assumption more closely by estimating the following semiparametric regressions of test scores on class size separately by each group k:

$$y_{isk} = \phi_k(n_{isk}) + x'_{isk}\theta_k + \varepsilon_{isk}\,. \tag{12}$$

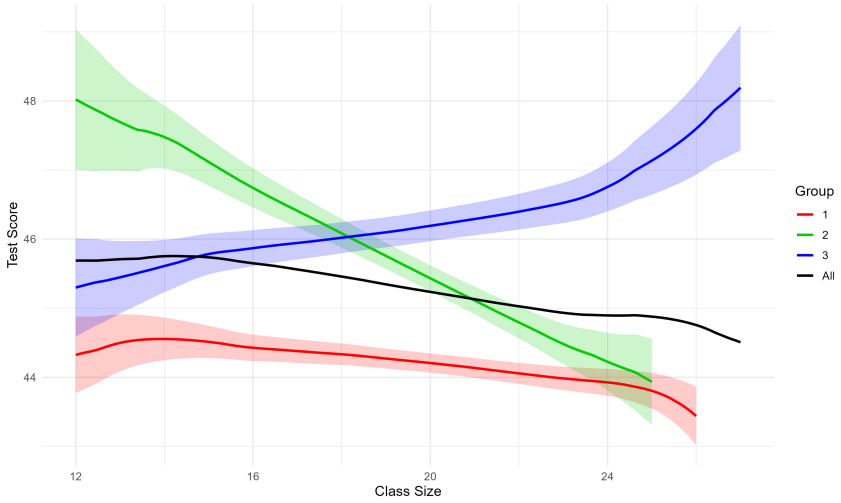
We estimate equation (12) using the Robinson (1988) semiparametric estimator, which first entails partialing out the group-specific means from both test scores and class sizes for the six cells created by $Female \times Nonwhite \times Freelunch$ before nonparametrically regressing demeaned test scores on demeaned class size using local linear regression. We use an Epanechnikov kernel and a bandwidth of 7.5 students, which lets in about 50% of the data.

Figure 3 displays the estimated regressions of test scores on class size over the support of class sizes within each group. The shaded regions represent asymptotic 95% confidence intervals. The sharp downward relationship between class size and test scores in group 2 is noteworthy. The difference in size of 12 students between the largest and smallest classes in group 2 produces a difference of approximately 1.25 sd on the

[15] We adjust for multiple hypothesis testing using p-values adjusted by the family-wise error rate over the six contrasts defined by race $\times$ group and free lunch status $\times$ group (Romano and Wolf 2005; Romano and Wolf 2016). Our bootstrapping procedure is designed to replicate the within-school randomization design by first sampling schools that participated in the STAR experiment and then sampling individuals within schools. The t-statistic for the hypothesis of no differential effects by group and race is 4.189 with a p-value for 0.036. The second-largest t-statistic is for differences in group one by free-lunch status ($t = 2.55$), but this is no longer significant after adjusting for multiple hypothesis testing $p = 0.363$.

[16] Bandiera, Larcinese, and Rasul (2010) find evidence of nonlinear effects of class size on academic performance among university students.

Figure 3: Nonparametric Regressions of Test Score on Class Size



The figure plots the semiparametric relationship between class size and test scores by group, as uncovered by the EAMP algorithm. All regressions use an Epanechnikov kernel and a bandwidth of 7.5. The regressions control for gender, race, and free lunch status. The black curve labeled *All* is a kernel-weighted average of the group-specific curves.

exam. This relationship dwarfs the effects seen in other groups. Although the slopes vary noticeably across groups, there is limited indication of nonlinearities within any of the groups. These nonparametric regressions additionally provide evidence of meaningful differences in slopes across the groups, reinforcing the finding in Table 2 that there is significant heterogeneity in the effectiveness of class size reductions across groups.

6 Ruling Out Spurious Heterogeneity

A potential concern with our approach is that, by clustering on ex post realizations of class sizes and test scores, the EAMP algorithm may overstate the extent of heterogeneity. If one clusters observations on outcomes, one will mechanically find clusters with different outcomes. Our approach guards against this concern in three ways.

6.1 The Importance of Model Selection Criteria

First, we use a model selection criteria to determine the number of groups. BIC strongly penalizes models with a large number of groups, because including an additional group raises the number of parameters to estimate by 21. This penalty ensures that the clusters represent distinct distributions of treatment effect heterogeneity rather than representing different regions of the same underlying distribution. If the distribution of class size

effects could have been rationalized by only one distribution, BIC would have selected one group.

We demonstrate this claim using a Monte Carlo simulation. Our particular simulation assumes no average effect of class size on test scores, but it allows for the possibility that the model is misspecified by including teacher effects, which are not present in the model in section 3, because STAR randomly assigned teachers to classes. We simulate data according to

$$y_{its} = \eta_s + \phi_t + \beta_{its} n_{its} + x'_{its}\theta + \varepsilon_{its} \ , \tag{13}$$

where ϕ_t represent teacher random effects and $\beta_{its} \sim N(0, 0.002)$. The school random effects are distributed as $\eta_s \sim N(0, 0.17)$, while the teacher random effects follow $\phi_t \sim N(0, 0.09)$.[17] We simulate data and estimate the model in section 3 95 times. We ask, across all 95 simulations, how often does BIC select a model with more than one group, in effect attributing teacher effects to class size effect heterogeneity?

BIC selects one cluster as optimal in all simulations. In no case does it interpret the presence of teacher heterogeneity across classrooms as evidence of differences in class size effectiveness across groups of schools. Even when choosing a model with multiple groups, the distribution of estimated group-specific means ($\mathbb{E}[\beta_{its} \mid S \in g$) is tightly centered around the truth. Figure C-1-(a) plots the density of the posterior school-specific mean effects in the Monte Carlo simulation for $K = 2$ groups. The figure shows that, even when applying the EAMP algorithm to estimate a misspecified model, the algorithm is still able to rationalize the data with a distribution of class size effects that is tightly centered around the truth. The distribution of school-specific posterior mean effects from the STAR data, plotted in Figure C-1-(b), instead has multiple modes. The difference between the Monte Carlo simulation and our empirical findings suggests that the sharp divergence in class size effects that we uncover in the STAR data is not due to model misspecification or overfitting.

[17] These numbers are chosen to be consistent with estimates of the standard deviation of teacher effects in Chetty, Friedman, and Rockoff (2014) and the standard deviation of school effects in (Angrist et al. 2024). Angrist et al. (2024) measure the standard deviation of school effects to be 0.19, but they do not try to separate the effects of teachers from the effects of schools. Under the assumption that the correlation between the two is zero, the standard deviation of $\eta_s + \phi_t$ given the numbers we choose equals the estimate of 0.19 from Angrist et al. (2024).

6.2 Groups Do Not Differ on Exogenous Characteristics Ex Post

A second concern that can arise from clustering on outcomes is that even if teachers were randomly assigned to classrooms ex ante, our model may have grouped schools together where good teachers happened to be assigned to smaller classrooms ex post. This would cause us to overstate the effectiveness of class size on test scores in the most effective group. If this were the case, the large effect of reducing class size on test scores seen in group 2 would not be due to class size itself, but due to schools in group 2 having better teachers in small classrooms due to random assignment. We believe this possibility can be discounted for several reasons. First, the previous section does not provide any evidence that our model displays this level of overfitting. Second, if the model had grouped schools together in this way, we may expect to see differences in observed teacher characteristics by classroom size within each group. But, Table C-8 shows that teacher characteristics are balanced across classroom types within each group.

Table 6 provides further evidence against this concern. It shows that group two schools produce large negative effects of class size on test scores in grades one, two, and three, while group three schools consistently have null to positive effects of class size on test scores in each grade. Random assignment of teachers to classrooms means that Kindergarten teacher quality should be orthogonal to teacher quality in subsequent grades. If the large class size effects at schools in group two had been caused by teacher quality, the effects would vanish in subsequent years. Despite substantial attrition from the original sample, Table 6 shows that the same schools consistently generate large effects of class size on test scores, suggesting that the gains are caused by class size itself and not by the presence of more skilled teachers in treated classrooms.

6.3 Effects on Additional Outcomes

A final concern related to ex post clustering is whether the treatment effect heterogeneity we recover can be generalized to other outcomes. We address this concern in two ways. First, we assess whether the same schools that generate large test score gains have impacts on later life outcomes. Second, we re-estimate our model with data on other outcomes measured in Kindergarten and ask how closely the findings for these other outcomes match our main findings.

Prior studies of the STAR experiment have shown that small Kindergarten classes raise the probability of taking a college entrance exam (Krueger and Whitmore 2001), the probability of graduating from high school (Finn, Gerber, and Boyd-Zaharias 2005), and attending college (Chetty et al. 2011). We examine whether these later-life outcomes were

Table 6: Class Size and Test Scores in Kindergarten through Third Grade

	Kindergarten		1st Grade		2nd Grade		3rd Grade	
Class size × Group 1	-0.064	-0.069	-0.026	-0.032	-0.082	-0.082	-0.094	-0.096
	(0.027)	(0.025)	(0.025)	(0.022)	(0.041)	(0.04)	(0.043)	(0.035)
Class size × Group 2	-0.305	-0.341	-0.206	-0.258	-0.194	-0.167	-0.174	-0.129
	(0.042)	(0.032)	(0.034)	(0.031)	(0.055)	(0.055)	(0.03)	(0.053)
Class size × Group 3	0.105	0.106	0.014	0.007	-0.054	-0.02	-0.042	-0.015
	(0.045)	(0.035)	(0.029)	(0.027)	(0.04)	(0.047)	(0.037)	(0.043)
Fixed Effects	Group	School	Group	School	Group	School	Group	School
N	3813	3813	2757	2757	2180	2180	1938	1909

The table shows the coefficients from OLS regressions of test scores in each grade on class size in the same grade with controls for nonwhite, free lunch status, and gender. All standard errors are clustered at the level of the school. The sample size decreases from left to right because of attrition.

driven by the same schools that raised Kindergarten test scores by estimating OLS regressions on class size interacted with group membership. Table 7 shows these regressions for several long-term outcomes. The first outcome, taking a college entrance exam (ACT or SAT), features no attrition. All students were matched with records from ACT, Inc and the College Board if they took either the ACT or SAT. The findings in columns (1) and (2) show that the entire effect of Kindergarten class size on exam-taking is driven by schools in group 2, the same schools where class size has a large effect on Kindergarten test scores. The estimates in column (2) show that a 1 student reduction in class size raises the probability of taking a college entrance exam by 1.2 percentage points at schools within group 2. Columns (3) and (4) show the same pattern for high school graduation, although the effects are only significant when including group, rather than school, fixed effects. Columns (5) and (6) show a similar story for high school GPA. Even though columns (3)-(6) feature substantial attrition and are not always significant, the estimates are consistent with this paper's main findings on the heterogeneity of class size effects across schools.

We additionally test whether class size causes changes in other outcomes measured at the end of Kindergarten. Our main outcome is a mean of math, reading, and word skill scores, but the STAR data features three additional skill measures in Kindergarten that we have not yet analyzed: listening scores, self concept scores, and motivation scores. The self-concept and motivation scores come from the Self-Concept and Motivation In-

Table 7: Effect of Kindergarten Class Size on Later Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	Take Exam		Graduate HS		HS GPA	
Class size × Group 1	0.002	0.001	−0.002	−0.004	−0.081	−0.052
	(0.003)	(0.003)	(0.003)	(0.003)	(0.080)	(0.069)
Class size × Group 2	−0.013***	−0.012**	−0.010*	−0.006	−0.126	−0.097
	(0.005)	(0.005)	(0.005)	(0.005)	(0.103)	(0.106)
Class size × Group 3	0.003	0.006**	0.004	0.007*	0.035	0.033
	(0.003)	(0.003)	(0.003)	(0.004)	(0.084)	(0.076)
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Group	School	Group	School	Group	School
N	3813	3813	1878	1878	1485	1485

Each column in the table displays estimates from OLS regressions on Kindergarten class size and controls for race, gender, and free lunch status where the coefficients on class size are allowed to vary by group membership. In columns (1)-(2) the dependent variable is a binary indicator for whether the student took a college entrance exam (SAT or ACT). In columns (3)-(4) the dependent variable is a binary indicator denoting graduation from high school. And in columns (5)-(6) the dependent variable is high school GPA. Table C-9 in Appendix C contains estimates for the demographic controls. All standard errors are clustered at the level of the school. *p<0.1; **p<0.05; ***p<0.01

ventory and use pictorial questions to measure children's achievement needs (the utility they derive from learning and social appreciation) and failure avoidance (Bietenbeck 2025). The listening assessment is a subtest of the Stanford Achievement Tests that measures oral language comprehension. The STAR experiment demonstrated a significant effect of attending a small class on the listening and self-concept scores but no effect on motivation scores (Word et al. 1990). For this reason, we re-estimate the model in section 3 using listening and self concept scores, respectively, as the outcome variable. These test scores are recorded for fewer numbers of students, so the sample size differs, which can affect the clustering of schools into groups based on the class size model. Nevertheless, Table E-3 shows that the estimated results for the listening score are qualitatively and quantitatively very similar to our main results. Group 2 has a large and significant effect of class size on the listening score, while the effects of class size are small and insignificant for the remaining groups. Table E-4 compares the grouping of schools in the main

outcome model and the listening model. The numbers on the diagonals are much larger than those off of the diagonals, indicating that most schools are assigned to the same group regardless of the outcome under study. This is noteworthy given the different outcome and the different sample size.

Table E-5 shows the model-estimated parameters for the self-concept score. Again, group 2 features a large effect of class size on self-concept scores, while no effects are seen for other groups. Table E-6 shows less similarity in the grouping of schools relative to the main outcome, but this is unsurprising given the much smaller sample size and the fact that the self-concept score is a measure of noncognitive skill. Altogether, the findings in this section reinforce our main conclusion of significant heterogeneity in the effectiveness of class size across schools, with one group generating large effects of class size across a range of outcomes.

7 Lessons for Scaling Class-Size Reduction Policies

Is it possible to use STAR to learn about how to implement effective class size reduction policies elsewhere? Prior literature has emphasized how supply-side factors, particularly a limited pool of high quality educators, pose a major challenge in scaling class size policies. For example, a state-wide reduction in class sizes in California in 1996 caused schools to hire less-qualified teachers, which contributed to the disappointing "voltage effects," whereby the policy's effects were much smaller than those observed in the experimental setting (List 2022).

Our paper provides a new framework for understanding the link between experimental interventions and widespread policy adoption. We argue that the first step in scaling an intervention is aligning the econometric model with the experimental design. Even after this is done, additional challenges remain, but we show how the GRE framework can provide tools to assess the likelihood that the intervention will scale in new contexts. We conclude by discussing when targeting resources may emerge as a useful scaling alternative to universal policies.

7.1 Assessing the Internal Validity of the STAR Evaluation

Following the discussion in section 2.2, we argue that the 2SLS estimator is not well-aligned with the STAR experimental design.[18] Table C-6 compares the results from our

[18]The same is true for the reduced-form estimator, as dose heterogeneity renders the estimates difficult to interpret.

Table 8: IV Weights

Group	Marginal Effect	ϕ_g	$P_{zg}(1-P_{zg})$	Dose Reduction	2SLS Weight
1	-0.068	0.374	0.249	-7.404	0.406
2	-0.339	0.298	0.247	-6.528	0.283
3	0.106	0.328	0.249	-6.465	0.311

The table displays instrumental variables weights computed according to equation (3). The columns labeled ϕ_g, $\overline{Z}_g(1-\overline{Z}_g)$ and dose reduction are sample estimates of the components of the 2SLS weight formula in equation (3).

estimated model with 2SLS that uses a treatment indicator as an instrument for class size in an IV model. The coefficient obtained by 2SLS (-0.101) does not differ much from the average marginal effect uncovered by the EAMP algorithm (-0.092), but this is more by chance than by design. Recall that in section 2.2 we showed how the 2SLS estimand equals

$$\beta_1^{2sls} = \sum_{s\in\mathcal{S}} \beta_s \frac{\phi_s P_{zs}(1-P_{zs})\left(\mathbb{E}[n_{is}|s, Z_{i,s}=1] - \mathbb{E}[n_{is}|s, Z_{i,s}=0]\right)}{\sum_{s\in\mathcal{S}} \phi_s P_{zs}(1-P_{zs})\left(\mathbb{E}[n_{is}|s, Z_{i,s}=1] - \mathbb{E}[n_{is}|s, Z_{i,s}=0]\right)},$$

where the school-specific weights on β_s depend on each schools' endogenous dose reduction. Table 8 aggregates these weights by group and breaks them down further into their individual components.[19] Schools in group 1 have the highest weight, because they created the largest dose reductions. Lower dose reductions cause schools in Groups 2 and 3 to be underweighted relative to their size. 2SLS does not differ significantly from the EAMP estimate, because endogenous compliance in STAR took a particular form that caused 2SLS to overweight the schools with the most typical impact effects (group 1) and underweight the schools in the tails of the distribution (groups 2 and 3).

Word et al. (1990) list two factors, teacher-identified discipline problems and parental complaints, that caused deviations from the STAR experimental design between Kindergarten and the first grade. It is conceivable that these concerns may have led some principals to target similar sizes for treatment and control classrooms at the program's inception, potentially leading schools with higher impact effects to implement smaller dose reductions.[20] An important takeaway from this episode is that interventions are often more likely to scale successfully when they receive approval and support from the local

[19] Appendix Figure C-2 plots the distribution of weights across schools within each group.

[20] School infrastructure may also have affected class size targets. The state of Tennessee covered the costs of hiring teachers and aides, but schools were required to provide additional classroom space if needed.

Table 9: Model Fit and Counterfactual Simulations

			Model's Counterfactual Simulation	
	Data	Model	Intended Experiment	Extreme Noncompliance
Reduced Form	0.727	0.645	0.762	1.902
2SLS	-0.101	-0.089	-0.090	-0.188
Avg. class size	18.991	19.02	19.764	20.131
S.D. class size	4.067	4.081	4.371	6.327
Dose Reduction	-7.228	-7.282	-8.417	-10.504

The table presents several moments from the actual data and simulations. The first row is the reduced form effect of randomization into a small class on test scores, estimated using a linear regression model with school fixed effects. The 2SLS model employs randomization into a small class as an instrument for class size in a two-stage least squares regression, incorporating school fixed effects. All regressions control for race, free lunch status, and gender. Simulated and counterfactual estimates are averages across 100 simulations. Details regarding the simulations are provided in section 7.

community (Agostinelli, Avitabile, and Bobba 2023). Such approval tends to be easier to obtain when the distribution of resources is perceived as equitable (Rawls 1971).

This may explain why all three groups of schools undercomplied with the STAR experimental protocols. To show this, we simulate the Tennessee STAR experiment as it was originally intended with small classes ranging from 13 to 17 students and regular classes between 22 and 25 students. We impose equal compliance by setting the Dirichlet distributions that govern class size to place equal weight on all sizes, so that $(\eta_k^{(t)}, \eta_k^{(c)}) = (\mathbf{1}, \mathbf{1})$ for all groups $k = 1, \ldots, K$. We then simulate class sizes and test scores and compute moments. These moments, in the third column of Table 9, show that the intended Tennessee STAR experiment would have generated a larger dose reduction, of over one student per classroom, between treated and control classrooms than was actually implemented. Endogenous compliance in STAR thus manifested itself in two ways: under-compliance overall and particularly large under-compliance at schools in Groups 2 and 3 that are associated with larger impact effects. If STAR had been implemented as intended, the reduced form effect would have been 18% higher (0.762 instead of 0.645).[21]

Alternative forms of noncompliance could have generated even larger effects. We sim-

[21]The counterfactual simulations in this section and in section 7.3 are ceteris paribus in the sense that they hold fixed other aspects of behavior that are not explicitly modeled.

ulate an implementation of STAR that maximizes the difference between treatment and control class sizes in Groups 1 and 2 and minimizes this difference for schools in group 3. This version of STAR would have generated an average difference in test scores between treated and control classrooms of ten and a half students, and a huge reduced form effect of 1.902 points (or 0.53 s.d.). The 2SLS estimate obtained from this experiment would be -0.188, which is twice the average marginal effect of -0.092 estimated by our model. This occurs because in this scenario 2SLS places no weight on schools in group 3, which have a positive effect of class size on test scores. Researchers relying on 2SLS using data from this hypothetical version of STAR would have erroneously concluded that reducing class size was twice as effective as it actually is.

These simulations show how the twin features of compliance and impact heterogeneity mean that different implementations of the STAR experiment on the same underlying population will generate different 2SLS estimates. While all estimates in Table 9 indicate a negative relationship between class size and test scores, the marginal effects estimated by 2SLS may differ considerably. These variable estimates can generate different rationales for action by policymakers, particularly when policies are justified on the basis of the marginal value of public funds (Mayshar 1990; Kline and Walters 2016). The 2SLS estimator is therefore poorly-aligned with an experimental design that allows for dose heterogeneity and endogenous dose responses. Furthermore, 2SLS tells researchers nothing about the heterogeneity in impact effects and compliance behavior, both of which affect scalability of the experimental intervention.

7.2 Assessing the Policy-Relevance of the STAR Evaluation for Scaling

GRE, by contrast, is robust to different implementations of the experiment and informs us about the distributions of compliance and impact heterogeneity. In the GRE framework, the policy-relevant treatment effect (PRTE) for moving from policy a to policy a' can be decomposed as follows:

$$PRTE_{a',a} = \sum_{k=1}^{K} \rho_k \mu_k \underbrace{\left(\mathbb{E}[n_{isk}|k, Policy = a'] - \mathbb{E}[n_{isk}|k, Policy = a]\right)}_{\text{Compliance}_k}, \tag{14}$$

Equation (14) shows that the expected effect of a policy intended to reduce class size hinges on schools' compliance with the policy, the structural impact effects μ_k, and the proportion of schools belonging to each group in the population, given by ρ_k. Al-Ubaydli et al. (2021) identify nonrepresentative populations and nonrepresentative behavioral

responses as two main threats to the scalability of experimental interventions. In our framework, ρ_k corresponds to representativeness of the population and $Compliance_k$ corresponds to representativeness of the behavioral response. An intervention may fail to scale because of changes in either ρ_k or $Compliance_k$.

The history of the Tennessee STAR experiment sheds light on the potential obstacles to implementing a statewide reduction in class size. Rockoff (2009) explains that around one in five eligible schools volunteered to participate. The type proportions, ρ_k, may therefore have differed between the experimental sample and the state of Tennessee. Schools that opted out of the experiment will likely be involved in any statewide class-size reduction policy, thereby causing the PRTE to differ from the average marginal effect identified in the experiment. However, by clustering schools based on their compliance and impact effect heterogeneity, GRE lets researchers model group membership as a function of site-specific covariates, in effect predicting ρ_k in new environments. The presence of significant predictors may hint at behavioral mechanisms that shape compliance or treatment effect heterogeneity, providing both new hypotheses for the researcher to test and clues for policymakers looking to scale the program effectively.

We demonstrate this by compiling a data set of school-specific predictors from STAR and estimating multinomial logit models of group membership given all possible permutations of these predictors. We vary the model size, steadily increasing the number of regressors, and use the BIC to select the optimal model. BIC selects a model with two regressors: the standard deviation of reading scores in the control classrooms and the fraction of teachers with a masters degree. The model indicates that a higher standard deviation of reading scores significantly predicts that a school belongs to groups 2 and 3, with the coefficient larger for group 3. A lower fraction of teachers with a masters degree positively predicts membership in group 3. Recall from Table 8 that groups two and three had smaller dose reductions than group 1. One potential explanation is that principals may be reluctant to allocate resources (class size) unequally in schools with greater inequality and fewer qualified teachers. More generally, this highlights that policymakers are not only concerned with identifying units with high marginal returns to treatment (such as schools in group 2), but also with selecting units that exhibit a high likelihood of complying with the policy itself.

Table 10 is a probabilistic mapping from site-specific covariates to group membership and provides a template for researchers who conduct small trials that they hope to scale. Researchers may use GRE to classify sites based on impact and compliance heterogeneity using data from the trial, estimating group membership as a function of site-specific

Table 10: Predictive Models

	Model 1		Model 2		Model 3	
	Group 2	Group 3	Group 2	Group 3	Group 2	Group 3
SD Reading: Control Group	9.31	17.43	10.78	19.39	8.66	13.94
	(4.40)	(4.72)	(4.69)	(5.11)	(5.50)	(5.91)
Teacher: Master's Degree			−1.89	−2.19	−1.92	−2.26
			(1.16)	(1.24)	(1.17)	(1.29)
Avg Letter Word: Control Group					1.72	4.43
					(2.40)	(2.55)
BIC	168.57		168.40		169.14	
Observations	79		79		79	

The table shows the estimated coefficients from three multinomial logit models predicting group membership as a function of school-specific covariates. Group one is the reference group. School-specific averages of individual- and teacher-specific variables are used. Standard errors are in parentheses.

covariates. If they subsequently collect data on the same covariates at new sites, they can use the multinomial model to predict where treatment effects and compliance will be greatest, in effect letting researchers forecast the PRTE in new populations by estimating the proportion of groups (ρ_k in equation 14). An additional advantage of this approach is that it allows for data-driven resource targeting in settings where researchers can accurately predict impact heterogeneity, allowing policy-makers to maximize program impacts for a fixed level of spending.

However, researchers and policymakers need to know how accurate their predictions will be before implementing targeted policies. We assess this by means of a leave-one-out predictive validity exercise, in which we predict each school's group membership by estimating Model 2 in Table 10 on the remaining 78 schools. We then sample the school's group membership from a multinomial model with $K = 3$ groups and probabilities given by the logit fitted values. Figure C-3 shows the fraction of times that this procedure assigns each school to the correct group. For groups 1 and 3, there is a wide distribution of predictive accuracy. Some schools are easy to assign to the correct group, while other schools are not. Group 2 instead has uniformly low predictive accuracy. Based on these findings, a policymaker should not have a great deal of confidence that they can accurately target resources to the schools (in group 2) where class size matters

the most.[22] This shows that, while GRE can be useful in identifying sites with greater impact effects or compliance, using STAR to learn were researchers should target class-size reductions is risky.

7.3 Targeting Small-Class Policies

Although STAR does not provide unambiguous advice for where to target class-size reductions, other interventions may be more informative. As an illustration, this section compares universal policies with targeted ones in which high-impact sites are known. We simulate a universal class-size reduction of five students across all schools and examine its effect on test scores and on the test-score gaps between Black and white students and between students who qualify for free lunch and those who do not. We then ask: What fraction of these effects could be achieved by a targeted policy that reduces class size by five students in the single group of schools with the largest marginal returns to class size?

Table 11 presents estimates of these counterfactual treatment effects. The first row shows that reducing class size by five for all students in the experiment would have raised test scores by 0.458 points overall, but the same intervention applied to only the 29% of schools in group 2 would have generated gains that are 10.3% larger. The second row shows that the universal reduction of five students would have lowered the Black-white test score gap by 0.137 points, but an intervention targeted only to group 2 schools would have reduced the gap by 0.105 points, or 76.7% of the original effect. The effects of both the universal and targeted interventions on the gap between students growing up in richer and poorer households (as measured by free lunch status) are small. These simulations suggest that there may be large cost efficiencies in policy implementation if resources can be targeted where they are likely to have the largest effects. Targeted policies can also be better monitored, making them less susceptible to a "voltage drop" in implementation.

8 Conclusion

This paper develops and implements a novel empirical framework that can be applied to large families of randomized controlled trials with potentially endogenous compliance

[22] A second reason to be cautious about extrapolating from STAR is the large variance in the random coefficient on class size in group 2 (0.235 in Table 2). While schools in group 2 have large class size effects on test scores on average, there is considerable variation across schools. Any policymaker interested in targeting resources to groups in school 2 might reasonably wonder which of the schools in the group are the most effective. We should note, however, that the sample size for this group is small and we do not have a high degree of confidence in the statistical accuracy of the random coefficient variance.

Table 11: Comparing Universal and Targeted Interventions

		Universal Intervention		Targeted Intervention	
	Data	Effect on Scores	Percent Reduction	Targeted Effect on Scores	Fraction of Universal Effect
Overall		0.458		0.506	1.103
Black-white gap	-1.395	0.137	9.8 %	0.105	0.767
Free lunch gap	-1.982	0.013	0.7 %	0.023	1.748

The table shows the effect of reducing class size by 5 students for everyone (universal intervention) with the effect of reducing class size by 5 students for schools in group two only (targeted intervention) on three separate outcomes. The first row simulates the effect on average test scores, while the second and third rows simulate the effects on the gaps in test scores between Black and white students and between students who qualify for free lunch and those who do not. Estimates are averages across 100 simulations. Details regarding the implementation of these counterfactual interventions are presented in section 7.

and heterogeneous treatment effects. We make use of the Grouped Random Effects approach of Adusumilli (2020), building on the Grouped Fixed Effects framework of Bonhomme and Manresa (2015), to identify patterns in treatment effect heterogeneity and compliance behavior across multi-site experiments. We then show how this framework can help researchers use the experimental data to learn about the scalability of the policy.

We apply this framework to data from the Tennessee STAR experiment to evaluate the causal effects of class size on student learning. The grouping algorithm uncovers marked differences in the effectiveness of class size reductions across schools. We find that 29% of schools in the experiment generated an average increase in test scores of 0.09 sd for each one-student reduction in class size. However, the effect of class size on test scores was small in the remaining schools.

Our findings contribute to the debate on the value of universal versus targeted interventions. We show how a well-targeted investment to reduce class size can generate similar overall effects on test scores as a universal program. The same program would also reduce the Black-white test score gap by 76.7% as much as a universal program. Targeted interventions may also have other benefits not studied in this paper, the most important being that that they can be scaled more easily while retaining quality.

By presenting evidence of greater heterogeneity in both impact effects and compliance behavior than previously acknowledged, we posit that the STAR experiment can reconcile contrasting conclusions from previous studies of class size effects on children's learning. We show how different implementations of STAR could have generated a wide range of estimates of class size effectiveness when using common regression estimators like TSLS. We additionally show that policy-relevant treatment effects (PRTEs) are affected by a particular setting's share of groups with different compliance and impact effect heterogeneity.

Our findings underscore the significance of designing pilot interventions capable of revealing the extent of treatment effect heterogeneity before advocating for the scaling of interventions across different contexts, situations, and populations. In addition, when experiments have high fixed costs and policymakers are impatient to scale results, List (2024) argues that individual experiments should generate information about the policy's scalability. We see the GRE framework as a tool to extract more information about treatment effect and compliance heterogeneity directly from a single experiment when it can be costly or infeasible to run additional ones. Political feasibility is also essential to effective scaling. Even when policymakers know where investments have the highest impact, targeting resources to only these locations can be either politically impossible or have deleterious long-run consequences (Weingast, Shepsle, and Johnsen 1981; Sen 1992; Burgess et al. 2015).

More research is still needed to understand why educational interventions cause such heterogeneous effects across schools. This paper provides a first step at documenting the wide variability in the production function relating schooling investments to knowledge. Our hope is that future researchers use these tools to better understand the economic and behavioral mechanisms that affect who responds endogenously to investments, who benefits from them, and why.

References

Achilles, C. M., J. B. Zaharias, B. A. Nye, and D. Fulton. 1995. "Analysis of Policy Application of Experimental Results: Project Challenge." Nasville, TN: Tennessee State University, Center of Excellence for Research in Basic Skills. Retrieved from http://eric.ed.gov/?id=ED393151.

Adusumilli, Karun. 2020. "Unobserved Heterogeneity, Grouped Random Effects and the EAMP Algorithm." *Unpublished Manuscript*.

Agostinelli, Francesco, Ciro Avitabile, and Matteo Bobba. 2023. "Enhancing Human Capital in Children: A Case Study on Scaling." Technical Report, National Bureau of Economic Research.

Al-Ubaydli, Omar, Min Sok Lee, John A List, Claire L Mackevicius, and Dana Suskind. 2021. "How Can Experiments Play a Greater Role in Public Policy? Twelve Proposals from an Economic Model of Scaling." *Behavioural Public Policy* 5 (1): 2–49.

Al-Ubaydli, Omar, John A. List, and Dana Suskind. 2020. "2017 Klein Lecture: The Science of Using Science: Toward an Understanding of the Threats to Scalability." *International Economic Review* 61 (4): 1387–1409.

Angrist, Joshua, Peter Hull, Parag A Pathak, and Christopher Walters. 2024. "Credible School Value-Added with Undersubscribed School Lotteries." *Review of Economics and Statistics* 106 (1): 1–19.

Angrist, Joshua D., and Guido W. Imbens. 1995. "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity." *Journal of the American Statistical Association* 90 (430): 431–442.

Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics* 114 (2): 533–575.

Angrist, Joshua D, Victor Lavy, Jetson Leder-Luis, and Adi Shany. 2019. "Maimonides Rule Redux." *American Economic Review: Insights* 1 (3): 309–324.

Bandiera, Oriana, Valentino Larcinese, and Imran Rasul. 2010. "Heterogeneous Class Size Effects: New Evidence from a Panel of University Students." *The Economic Journal* 120 (549): 1365–1398.

Bergman, Peter, Raj Chetty, Stefanie DeLuca, Nathaniel Hendren, Lawrence F Katz, and Christopher Palmer. 2024. "Creating Moves to Opportunity: Experimental Evidence on Barriers to Neighborhood Choice." *American Economic Review* 114 (5): 1281–1337.

Bietenbeck, Jan. 2025. "Do Motivated Classmates Matter for Educational Success?" *The Economic Journal* 135 (665): 36–58.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe. 2017. "Variational Inference: A Review for Statisticians." *Journal of the American statistical Association* 112 (518): 859–877.

Bohrnstedt, George W, and Brian M Stecher. 1999. "Class Size Reduction in California 1996-1998: Early Findings Signal Promise and Concerns. Report Summary."

———. 2002. "What We Have Learned about Class Size Reduction in California. Capstone Report."

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2022. "Discretizing Unobserved Heterogeneity." *Econometrica* 90 (2): 625–643.

Bonhomme, Stéphane, and Elena Manresa. 2015. "Grouped Patterns of Heterogeneity in Panel Data." *Econometrica* 83 (3): 1147–1184.

Borghesan, Emilio, and Gabrielle Vasey. 2024. "The Marginal Returns to Distance Education: Evidence from Mexico's Telesecundarias." *American Economic Journal: Applied Economics* 16 (1): 253–285.

Boyd-Zaharias, J, J Finn, R Fish, and S Gerber. 2007. "Project STAR and Beyond: Database User's Guide." *HEROS Inc. and University of New York at Buffalo technical report*.

Burgess, Robin, Remi Jedwab, Edward Miguel, Ameet Morjaria, and Gerard Padró i Miquel. 2015. "The Value of Democracy: Evidence from Road Building in Kenya." *American Economic Review* 105 (6): 1817–1851.

Card, David, and Alan B. Krueger. 1992a. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy* 100 (1): 1–40.

———. 1992b. "School Quality and Black-White Relative Earnings: A Direct Assessment." *The Quarterly Journal of Economics* 107 (1): 151–200.

Carneiro, Pedro, James J Heckman, and Edward J Vytlacil. 2011. "Estimating Marginal Returns to Education." *American Economic Review* 101 (6): 2754–2781.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *The Quarterly Journal of Economics* 126 (4): 1593–1660.

Chetty, Raj, John N Friedman, and Jonah E Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American economic review* 104 (9): 2593–2632.

de Chaisemartin, Clément, and Antoine Deeb. 2024. "Estimating Treatment-effect Heterogeneity across Sites in Multi-site Randomized Experiments with Imperfect Compliance." *arXiv preprint arXiv:2405.17254*.

DellaVigna, Stefano, Woojin Kim, and Elizabeth Linos. 2024. "Bottlenecks for Evidence Adoption." *Journal of Political Economy* 132 (8): 2748–2789.

Ding, Weili, and Steven F Lehrer. 2010. "Estimating Treatment Effects from Contaminated Multiperiod Education Experiments: The Dynamic Impacts of Class Size Reductions." *The Review of Economics and Statistics* 92 (1): 31–42.

Finn, Jeremy D, and Charles M Achilles. 1990. "Answers and Questions about Class Size: A Statewide Experiment." *American Educational Research Journal* 27 (3): 557–577.

Finn, Jeremy D, Susan B Gerber, and Jayne Boyd-Zaharias. 2005. "Small Classes in the Early Grades, Academic Achievement, and Graduating from High School." *Journal of Educational Psychology* 97 (2): 214.

Folger, John, and Carolyn Breda. 1989. "Evidence from Project STAR about Class Size and Student Achievement." *Peabody Journal of Education* 67 (1): 17–33.

Gechter, Michael, Keisuke Hirano, Jean Lee, Mahreen Mahmud, Orville Mondal, Jonathan Morduch, Saravana Ravindran, and Abu S Shonchoy. 2023. "Site Selection for External Validity: Theory and an Application to Mobile Money in South Asia."

Hanushek, Eric A. 1997. "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19 (2): 141–164.

Heckman, James. 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs. Edited by C. F. Manski and I. Garfinkel*. Harvard University Press.

Heckman, James, Anne Layne-Farrar, and Petra Todd. 1995, September. "Does Measured School Quality Really Matter? An Examination of the Earnings-Quality Relationship." Working paper 5274, National Bureau of Economic Research.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics* 88 (3): 389–432.

Heckman, James J, and Edward J Vytlacil. 2001. "Policy-Relevant Treatment Effects." *American Economic Review* 91 (2): 107–111.

Hippel, Paul von, and Chandi Wagner. 2018. "Does a Successful Randomized Experiment Lead to Successful Policy? Project Challenge and What Happened in Tennessee After Project STAR (March 31, 2018)." Available at SSRN: https://ssrn.com/abstract=3153503.

Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *The Quarterly Journal of Economics* 115 (4): 1239–1285 (11).

Jepsen, Christopher, and Steven Rivkin. 2009. "Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size." *The Journal of Human Resources* 44 (1): 223–250.

Kline, Patrick, Evan K Rose, and Christopher R Walters. 2022. "Systemic Discrimination among Large US Employers." *The Quarterly Journal of Economics* 137 (4): 1963–2036.

Kline, Patrick, and Christopher Walters. 2021. "Reasonable Doubt: Experimental Detection of Job-Level Employment Discrimination." *Econometrica* 89 (2): 765–792.

Kline, Patrick, and Christopher R Walters. 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start." *The Quarterly Journal of Economics* 131 (4): 1795–1848.

Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *The Quarterly Journal of Economics* 114 (2): 497–532.

Krueger, Alan B, and Diane M Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College-test Taking and Middle School Test Results: Evidence from Project STAR." *The Economic Journal* 111 (468): 1–28.

List, John A. 2022. *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. Currency.

———. 2024. "Optimally Generate Policy-Based Evidence before Scaling." *Nature* 626 (7999): 491–499.

Mayshar, Joram. 1990. "On Measures of Excess Burden and Their Application." *Journal of Public Economics* 43 (3): 263–289.

Mishel, Lawrence, and Richard Rothstein. 2002. In *The Class Size Debate*, 1–102. Economic Policy Institute.

Neal, Radford M, and Geoffrey E Hinton. 1998. "A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants." In *Learning in Graphical Models*, 355–368. Springer.

Raudenbush, Stephen W, and Howard S Bloom. 2015. "Learning about and from a Distribution of Program Impacts Using Multisite Trials." *American Journal of Evaluation* 36 (4): 475–499.

Rawls, John. 1971. *A Theory of Justice: Original Edition*. Harvard University Press.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–458.

Robinson, Peter M. 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica: Journal of the Econometric Society*, pp. 931–954.

Rockoff, Jonah. 2009. "Field Experiments in Class Size from the Early Twentieth Century." *Journal of Economic Perspectives* 23 (4): 211–30 (December).

Romano, Joseph P, and Michael Wolf. 2005. "Stepwise Multiple testing as Formalized Data Snooping." *Econometrica* 73 (4): 1237–1282.

———. 2016. "Efficient Computation of Adjusted P-values for Resampling-Based Stepdown Multiple Testing." *Statistics & Probability Letters* 113:38–40.

Rose, Evan K., and Yotam Shem-Tov. 2021. "How Does Incarceration Affect Reoffending? Estimating the Dose-Response Function." *Journal of Political Economy* 129 (12): 3302–3356.

Schanzenbach, Diane Whitmore. 2006. "What Have Researchers Learned from Project STAR?" *Brookings Papers on Education Policy*, no. 9:205–228.

———. 2016. *Long-term Impacts of Class Size Reduction. In Peter Blatchford, et al., eds., International Perspectives on Class Size*. Routledge.

Sen, Amartya. 1992. *The Political Economy of Targeting*. World Bank Washington, DC.

Stecher, Brian M, and George W Bohrnstedt. 2000. "Class Size Reduction in California: Summary of the 1998-99 Evaluation Findings."

Walters, Christopher R. 2018. "The Demand for Effective Charter Schools." *Journal of Political Economy* 126 (6): 2179–2223.

Weingast, Barry R, Kenneth A Shepsle, and Christopher Johnsen. 1981. "The Political Economy of Benefits and Costs: A Neoclassical Approach to Distributive Politics." *Journal of Political Economy* 89 (4): 642–664.

Word, Elizabeth, John Johnston, Helen P Bain, B DeWayne Fulton, Jayne B Zaharias, Charles M Achilles, Martha N Lintz, John Folger, and Carolyn Breda. 1990. "The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project." *Tennessee Board of Education*.

Online Appendix

A Details of the EAMP algorithm

This appendix describes the steps of the EAMP algorithm used to estimate the GRE model.

Step E: Expectation

By the Donsker-Varadhan variational formula, the optimal value of $q_{sk}(\boldsymbol{\alpha}_s)$ is just the posterior distribution of $\boldsymbol{\alpha}_s$, as implied by the likelihood $p(\boldsymbol{y}_s, \boldsymbol{n}_s^{(c)}, \boldsymbol{n}_s^{(t)} | \boldsymbol{\alpha}_s, \boldsymbol{x}, \boldsymbol{\sigma}_\epsilon^2, \theta)$ and the prior $\pi(\boldsymbol{\alpha}_s | \boldsymbol{\rho}_k)$. Since the prior is conjugate to the likelihood, the posterior can be computed very quickly. To characterize the posterior, we first note that due to the structure of the model, the posterior is separable:

$$q_{sk}(\boldsymbol{\alpha}_s) = q_{sk}(\boldsymbol{\beta}_s) \cdot q_{sk}(\boldsymbol{p}_s^{(c)}) \cdot q_{sk}(\boldsymbol{p}_s^{(t)}).$$

We can then update each of these quantities separately as follows. The update to the posterior distribution of $\boldsymbol{\beta}_s$ is given by

$$q_{sk}(\boldsymbol{\beta}_s) \equiv \prod_{i:s(i)=s} q_{sk}(\beta_{isk}) \leftarrow \prod_{i:s(i)=s} N(\beta_{isk} | \mu_{isk}, \Sigma_{isk})$$

where, for each i such that $s(i) = s$, we update

$$\begin{aligned} \Sigma_{isk} &\leftarrow \left(\Sigma_k^{-1} + \frac{n_{isk} n_{isk}^\intercal}{\sigma_{\epsilon,k}^2} \right)^{-1}, \\ \mu_{isk} &\leftarrow \Sigma_{isk} \left(\Sigma_k^{-1} \mu_k + \frac{n_{isk}(y_{isk} - x'_{isk}\theta)}{\sigma_{\epsilon,k}^2} \right) \end{aligned} \tag{A-1}$$

The update to $q_{sk}(p_s^{(c)})$ is given by

$$q_{sk}(p_{sk}^{(c)}) \leftarrow \text{Dirichlet}(p_{sk}^{(c)} | \eta_k^{(c)}),$$

where $\eta_k^{(c)} \equiv \{\eta_{k1}^{(c)}, \ldots, \eta_{kL}^{(c)}\}$ denotes the posterior values of $\eta_k^{(c)}$, and is given by

$$\eta_{kl}^{(c)} = \eta_{kl} + \sum_{g=1}^{N^{(c)}(s)} \mathbb{I}\{n_{skg}^{(c)} \equiv 16 + m\}, \text{ for each } m = 0, \ldots, M.$$

The update to $q_{sk}(p_{sk}^{(t)})$ is analogous.

Step A: Assignment

We assign each observation to one of the K groups by maximizing (11) with respect to $w_s(k)$. As in Adusumilli (2020), group assignments are obtained as the solution to the following problem:

$$k(s) \leftarrow \arg\max_k I_{sk};\ I_{sk} := \ln \frac{p(\boldsymbol{y}_s, \boldsymbol{n}_s^{(c)}, \boldsymbol{n}_s^{(t)} | \boldsymbol{\alpha}_s, \boldsymbol{x}, \boldsymbol{\sigma}_\epsilon^2, \theta) \pi\left(\boldsymbol{\alpha}_s | \boldsymbol{\rho}_k\right)}{q_{sk}(\boldsymbol{\alpha}_s)}. \tag{A-2}$$

Since the posterior distribution, $q_{sk}(\boldsymbol{\alpha}_s)$, is known from Step E, we can obtain an analytical expression for I_{sk}. To obtain this expression, note that

$$\ln I_{sk} = \frac{p(\boldsymbol{y}_s | \boldsymbol{\beta}_s, \boldsymbol{x}, \sigma^2_{\epsilon,k(s)}, \theta) \cdot p(\boldsymbol{n}_s^{(c)} | \boldsymbol{p}_s^{(c)}) \cdot p(\boldsymbol{n}_s^{(t)} | \boldsymbol{p}_s^{(t)}) \pi(\boldsymbol{\beta}_s | \boldsymbol{\rho_k}) \cdot \pi(\boldsymbol{p}_s^{(c)} | \boldsymbol{\eta}_k^{(c)}) \cdot \pi(\boldsymbol{p}_s^{(t)} | \boldsymbol{\eta}_k^{(t)})}{q_{sk}(\boldsymbol{\beta}_s) \cdot q_{sk}(\boldsymbol{p}_s^{(c)}) \cdot q_{sk}(\boldsymbol{p}_s^{(t)})}, \tag{A-3}$$

which can be grouped into three separate terms:

$$\begin{aligned} \ln \frac{p(\boldsymbol{y}_s | \boldsymbol{\beta}_s, \boldsymbol{x}, \boldsymbol{\sigma}^2_{\epsilon,k(s)}, \theta) \pi(\boldsymbol{\beta}_s | \boldsymbol{\rho_k})}{q_{sk}(\boldsymbol{\beta}_s)} &= -\frac{1}{2} \sum_{i:s(i)=s} \left\{ \frac{(y_{isk} - x'_{isk}\theta)^2}{\sigma^2_{\epsilon,k}} + \mu_k^\top \Sigma_k^{-1} \mu_k - \mu_{ksi}^\top \Sigma_{ksi}^{-1} \mu_{ksi} \right\} \\ &\quad + \frac{1}{2} \sum_{i:s(i)=s} \{\ln |\Sigma_{ksi}| - \ln |\Sigma_k|\} - \frac{1}{2} \ln \sigma^2_{\epsilon,k} + \text{ const} \end{aligned} \tag{A-4}$$

$$\begin{aligned} \ln \frac{p(\boldsymbol{n}_s^{(c)} | \boldsymbol{p}_s^{(c)}) \pi(\boldsymbol{p}_s^{(c)} | \boldsymbol{\eta}_k^{(c)})}{q_{sk}(\boldsymbol{p}_s^{(c)})} &= \ln \frac{N_s^{(c)}!}{\prod_{j=1}^J n_{skj}^{(c)}!} \frac{B(\eta_k^{(c)} + n_{sk}^{(c)})}{B(\eta_k^{(c)})} \\ &= \ln B(\eta_k^{(c)} + n_{sk}^{(c)}) - \ln B(\eta_k^{(c)}) + \text{const} \end{aligned} \tag{A-5}$$

$$\begin{aligned} \ln \frac{p(\boldsymbol{n}_s^{(t)} | \boldsymbol{p}_s^{(t)}) \pi(\boldsymbol{p}_s^{(t)} | \boldsymbol{\eta}_k^{(t)})}{q_{sk}(\boldsymbol{p}_s^{(t)})} &= \ln \frac{N_s^{(t)}!}{\prod_{j=1}^J n_{skj}^{(t)}!} \frac{B(\eta_k^{(t)} + n_{sk}^{(t)})}{B(\eta_k^{(t)})} \\ &= \ln B(\eta_k^{(t)} + n_{sk}^{(t)}) - \ln B(\eta_k^{(t)}) + \text{const} \end{aligned} \tag{A-6}$$

where $n_{sk}^{(c)}$, and $n_{sk}^{(t)}$ are the count vectors containing the number of classes of each type in school s, $N_s^{(c)}$ and $N_s^{(t)}$ are the number of control and treatment classrooms in school s, $B(\eta) = \frac{\prod_{j=1}^J \Gamma(\eta_j)}{\Gamma(\sum_{j=1}^J \eta_j)}$, and $\Gamma(n) = (n-1)!$. We compute I_{sk} as the sum of (A-4) (A-5) and (A-6) separately for each school and group, and then we assign each school to the group with the greatest value of I_{sk}.

Step M: Maximization

The maximization step updates the estimates of the nonrandom parameters, θ and $\boldsymbol{\sigma}_{\epsilon}^2$. We compute θ by solving

$$\begin{aligned}
&\max_{\theta} \sum_{s=1}^{S}\sum_{k=1}^{K} w_s(k) E_{q_{sk}(\cdot)}\left[\ln p(\boldsymbol{y}_s, \boldsymbol{n}_s^{(c)}, \boldsymbol{n}_s^{(t)}|\boldsymbol{\alpha}_s, \boldsymbol{x}, \sigma_{\epsilon,k}^2, \theta)\right] = \\
&\max_{\theta} \sum_{s=1}^{S}\sum_{k=1}^{K} w_s(k) E_{q_{sk}(\cdot)}\left[\ln p(\boldsymbol{y}_s|\boldsymbol{\alpha}_s, \boldsymbol{x}, \sigma_{\epsilon,k}^2, \theta) + \ln p(\boldsymbol{n}_s^{(t)}|\boldsymbol{\alpha}_s) + \ln p(\boldsymbol{n}_s^{(c)}|\boldsymbol{\alpha}_s)\right] = \\
&\max_{\theta} \sum_{s=1}^{S}\sum_{k=1}^{K} w_s(k) E_{q_{sk}(\cdot)}\left[\ln p(\boldsymbol{y}_s|, \boldsymbol{\alpha}_s, \boldsymbol{x}, \sigma_{\epsilon,k}^2, \theta)\right] \qquad \text{(A-7)}
\end{aligned}$$

Since y_{isk} is normally distributed conditional on covariates, class size, and the random coefficients, the solution to (A-7) is a linear projection:

$$\theta = \left(\sum_{i=1}^{N} x_{isk} x_{isk}'\right)^{-1} \left(\sum_{i=1}^{N}(y_{isk} - \sum_{k=1}^{K} w_{s(i)}(k) n_{isk} E_{q_{sk}(\cdot)}[\beta_{isk}])\right) , \qquad \text{(A-8)}$$

where $w_{s(i)}(k)$ is an indicator for whether student i's school (s) belongs to group k, and $E_{q_{sk}(\cdot)}[\beta_{isk}]$ is the posterior mean of β_{isk}, specifically μ_{isk} from equation (A-1). Note that because we are taking expectation with respect to the posterior distribution conditional on observing the data (including class size), class size is in the conditioning set so that

$$E_{q_{sk}(\cdot)}[w_{s(i)}(k) n_{isk} \beta_{isk}] = w_{s(i)}(k) n_{isk} E_{q_{sk}(\cdot)}[\beta_{isk}] ,$$

which delivers the formula in (A-8).

We then compute $\sigma_{\epsilon,k}^2$ for $k = 1, \ldots, K$ by solving

$$\begin{aligned}
\sigma_{\epsilon,k}^2 &= \frac{1}{n_k} \sum_{i=1}^{N} w_{s(i)}(k) E_{q_{isk}(\cdot)}\left[(y_{isk} - \beta_{isk} n_{isk} - x_{isk}'\theta)^2\right] , \\
&= \frac{1}{n_k} \sum_{i=1}^{N} w_{s(i)}(k) \left[(y_{isk} - \mu_{isk} n_{isk} - x_{isk}'\theta)^2 + \Sigma_{isk} n_{isk}^2\right] ,
\end{aligned}$$

where n_k is the number of students in group k, and $E_{q_{isk}}(\beta_{isk})$ and $Var(\beta_{isk})$ are posterior means and variances computed in (A-1).

Step P: Propagation

The prior is from the exponential family. Hence, as in Adusumilli (2020), updating the prior parameters involves matching the sufficient statistics of the exponential family between the prior and average posterior. Due to separability of both the prior and posterior, we can separately update the prior parameters γ_k, $\eta_k^{(c)}$, and $\eta_k^{(t)}$.

We update the mean and variance for each group as follows:

$$\mu_k \leftarrow \frac{1}{n_k}\sum_s w_s(k) \sum_{i:s(i)=s} \mu_{isk}, \tag{A-9}$$

$$\Sigma_k \leftarrow \frac{1}{n_k}\sum_s w_s(k) \sum_{i:s(i)=s} \left\{\Sigma_{isk} + \mu_{isk}\mu_{isk}^{\intercal}\right\} - \mu_k\mu_k^{\intercal}, \tag{A-10}$$

where n_k is the number of observations (students) in group k. If, in the process of optimization, a group turns out to be empty, we do not update the posterior for that group.

To update $\eta_k^{(c)}$, we match the posterior average and prior moments of $\ln p_{skl}^{(c)}$ for each l as these are the sufficient statistics of the Dirichlet family. This implies that $\eta_k^{(c)} \equiv (\eta_{k1}^{(c)}, \ldots, \eta_{kL}^{(c)})$ can be obtained as the solution to the system of L equations Denote by $\tilde{\eta_{kj}}^{(c)}$ the updated parameter for control class type j in group k. Then the system of L equations in L unknowns for control group k is given by

$$\psi\left(\tilde{\eta}_{kl}^{(c)}\right) - \psi\left(\sum_{l=1}^{L}\tilde{\eta}_{kl}^{(c)}\right) = \frac{1}{N_{sk}}\sum_s w_s(k)\left\{\psi\left(\eta_{skl}^{(c)}\right) - \psi\left(\sum_{l=1}^{L}\eta_{skl}^{(c)}\right)\right\}, \text{ for each } l = 1, \ldots, L \tag{A-11}$$

where $\psi(\cdot)$ denotes the Digamma function, N_{sk} is the number of schools in group k, and the right hand side variables, $\eta_{sk}^{(c)}$, where obtained in the E-step as the sum of the Dirichlet prior and the vector of class size counts for school s: $\eta_{sk}^{(c)} = \eta_k^{(c)} + \boldsymbol{n}_{sk}$.

The system of equations for treatment group k is analogous:

$$\psi\left(\tilde{\eta}_{kl}^{(t)}\right) - \psi\left(\sum_{m=1}^{M}\tilde{\eta}_{km}^{(t)}\right) = \frac{1}{N_{sk}}\sum_s w_s(k)\left\{\psi\left(\eta_{skm}^{(t)}\right) - \psi\left(\sum_{m=1}^{M}\eta_{skm}^{(t)}\right)\right\}, \text{ for each } m = 1, \ldots, M \tag{A-12}$$

Note that, because the test score model generates scores for students, but the class size model generates counts of classes, N_{sk} in equations (A-11) and (A-12) refer to the number of schools.

These systems of equations are solved for each group and treatment/control status to obtain $2XK$ posterior parameter vectors, $\tilde{\eta}_k^{(c)}$ and $\tilde{\eta}_k^{(t)}$ for each group, $k = 1, \ldots, K$.

B Proof of Proposition 1

We first demean the following linear model relating class size to test scores

$$y_{is} = \eta_s + \beta_{is} n_{is} + \epsilon_{is}$$

by their population means to obtain

$$\underbrace{y_{is} - \mathbb{E}[y_{is} \mid s]}_{\Delta y_{is}} = \beta_{is} \underbrace{(n_{is} - \mathbb{E}[n_{is} \mid s])}_{\Delta n_{is}} + \underbrace{(\epsilon_{is} - \mathbb{E}[\epsilon_{is} \mid s])}_{\Delta \epsilon_{is}}$$

The 2SLS estimand for this demeaned model is:

$$\begin{aligned}
\beta_1^{2SLS} &= \frac{\mathbb{E}[\Delta y_{is} \Delta z_{is}]}{\mathbb{E}[\Delta n_{is} \Delta z_{is}]} \\
&= \frac{\mathbb{E}[(\beta_{is} \Delta n_{is} + \Delta \epsilon_{is}) \Delta z_{is}]}{\mathbb{E}[\Delta n_{is} \Delta z_{is}]} \\
&= \frac{\mathbb{E}[\beta_{is} \Delta n_{is} \Delta z_{is}]}{\mathbb{E}[\Delta n_{is} \Delta z_{is}]} \\
&= \frac{\mathbb{E}[\mathbb{E}[\beta_{is} \Delta n_{is} \Delta z_{is} \mid s]]}{\mathbb{E}[\Delta n_{is} \Delta z_{is}]} \\
&= \frac{\mathbb{E}[\beta_s \mathbb{E}[\Delta n_{is} \Delta z_{is} \mid s]]}{\mathbb{E}[\Delta n_{is} \Delta z_{is}]} \\
&= \sum_s \frac{\phi_s \beta_s \mathbb{E}(z_{is} \mid s)(1 - \mathbb{E}(z_{is} \mid s)) \left(\mathbb{E}[n_{is} \mid s, z_{is} = 1] - \mathbb{E}[n_{is} \mid s, z_{is} = 0]\right)}{\sum_s \phi_s \mathbb{E}(z_{is} \mid s)(1 - \mathbb{E}(z_{is} \mid s)) \left(\mathbb{E}[n_{is} \mid s, z_{is} = 1] - \mathbb{E}[n_{is} \mid s, z_{is} = 0]\right)}
\end{aligned}$$

where ϕ_s in the final expression is the population fraction of students enrolled in school s. The fifth equality follows from random assignment of students to classrooms, which ensures that the random coefficients are independent of class size conditional on the school, and the sixth inequality follows from the binary nature of z_{is}:

$$\begin{aligned}
\mathbb{E}[\Delta n_{is} \Delta z_{is} \mid s] =& \mathbb{E}[\Delta n_{is}(1 - \mathbb{E}[z_{is} \mid s]) \mathbb{1}_{z_{is}=1} \mid s] - \mathbb{E}[\Delta n_{is} \mathbb{E}[z_{is} \mid s] \mathbb{1}_{z_{is}=0} \mid s] \\
=& \mathbb{E}[\Delta n_{is}(1 - \mathbb{E}[z_{is} \mid s]) \mid s, z_{is} = 1] \mathbb{E}[z_{is} \mid s] - \\
& \mathbb{E}[\Delta n_{is} \mathbb{E}[z_{is} \mid s] \mid s, z_{is} = 0](1 - \mathbb{E}[z_{is} \mid s]) \\
=& \mathbb{E}[z_{is} \mid s](1 - \mathbb{E}[z_{is} \mid s]) \left(\mathbb{E}[\Delta n_{is} \mid s, z_{is} = 1] - \mathbb{E}[\Delta n_{is} \mid s, z_{is} = 0]\right) \\
=& \mathbb{E}[z_{is} \mid s](1 - \mathbb{E}[z_{is} \mid s]) \left(\mathbb{E}[n_{is} \mid s, z_{is} = 1] - \mathbb{E}[n_{is} \mid s, z_{is} = 0]\right)
\end{aligned}$$

where the last equality results from taking $\mathbb{E}[n_{is} \mid s]$ out of the two conditional expectations.

C Additional Tables and Figures - Main Sample

Table C-1: Summary Statistics

	Mean	Standard Deviation
Test Score	45.30	3.586
Class Size	18.991	4.067
Female	0.489	0.5
Not White/Asian	0.319	0.466
Eligible for Free Lunch	0.473	0.499
Treated	0.465	0.499

The table presents descriptive statistics for the sample of 3813 school children used in estimation. Test Score is the average of student scores on the math, reading, and word skills SAT exams administered at the end of kindergarten. Female, Nonwhite, and Eligible for Free Lunch are all binary variables.

Table C-2: Student Characteristics by Group

Group	Female	Nonwhite	Free lunch	Avg. Days Present	S.D. Days Present
1	0.48	0.32	0.48	158.0	25.7
2	0.49	0.37	0.48	153.4	27.6
3	0.50	0.30	0.48	157.3	24.6

The table shows average student characteristics for students attending schools in each group. Avg. Days Present is determined by both the length of the school year and student absences.

Table C-3: Teacher Characteristics by Group

Group	Masters Degree	Experience	Nonwhite
1	0.42	8.80	0.17
2	0.31	9.52	0.17
3	0.31	9.49	0.14

The table shows average teacher characteristics for teachers at schools in each group. Experience is measured in years.

Table C-4: School Characteristics by Group

Group	Mean Class Size	S.D. Class Size	Cohort Size	S.D. Cohort Size	% Race-Segregated	% SES-Segregated
1	20.92	4.23	78.5	19.3	0.81	0.29
2	20.04	3.52	82.3	29.1	0.96	0.39
3	19.92	4.00	79.9	31.3	0.92	0.36

Group	Inner City	Rural	Suburban	Urban
1	8	13	6	4
2	5	10	6	2
3	3	15	6	1

The table shows school characteristics by group membership. A school is classified as race-segregated if over 80% of the student body belongs to a single racial or ethnic group. A school is classified as SES-segregated if either over 80% or under 20% of the student body qualifies for free or reduced price lunch.

Table C-5: Estimates of the Model for Test Scores

Group	μ_k	Σ_k	$\sigma^2_{\epsilon,k}$	Schools	Students
1	-0.068	0.042	4.933	31	1425
	(0.061)	(0.065)	(1.55)		
2	-0.339	0.235	6.442	23	1137
	(0.062)	(0.15)	(1.86)		
3	0.106	0.001	11.852	25	1251
	(0.059)	(0.07)	(2.07)		
Covariates					
Female	0.666				
	(0.14)				
Nonwhite	-1.109				
	(0.29)				
Free Lunch	-1.683				
	(0.20)				
N	3813				

The table shows the estimated model parameters governing the effect of class size on test scores including controls for gender, race, and free lunch status. Controls are constrained to be equal across groups. The class size effects represent the effect of a one-unit increase in class size on test score performance. Bootstrapped standard errors from 95 bootstrapped data sets are in parentheses. Bootstrapping involves first sampling schools from the set of 79 schools that participated in the experiment and then sampling individuals within each school.

Table C-6: 2SLS Model

	2SLS
Class Size	-0.101
	(0.014)
Female	0.650
	(0.099)
Nonwhite	-1.057
	(0.209)
Free Lunch	-1.776
	(0.119)
First Stage F-Statistic	42237.32
Observations	3823

The table presents estimates from a two-stage least squares regression that uses the binary randomization into a small classroom as the excluded instrument. Heteroskedasticity-robust standard errors are in parentheses.

Table C-7: Model Fit

	All		Group 1		Group 2		Group 3	
	Data	Sim	Data	Sim	Data	Sim	Data	Sim
Avg Score	45.300	45.308	44.227	44.242	45.707	45.685	46.153	46.171
SD Score	3.586	6.612	2.856	4.957	3.710	9.850	3.902	3.940
Avg Class Size	18.991	19.02	19.359	19.388	18.777	18.81	18.765	18.789
SD Class Size	4.067	4.081	4.325	4.338	3.616	3.590	4.124	4.146

The table compares moments of the model's endogenous variables, class size and test scores, overall and by group to their counterparts in the data.

Table C-8: Within-group Teacher Characteristics by Treatment Assignment

	Black			Master's Degree			Experience		
Group	Control	Treat	p-value	Control	Treat	p-value	Control	Treat	p-value
1	0.206	0.146	0.494	0.429	0.396	0.769	8.371	9.021	0.599
2	0.194	0.194	0.993	0.258	0.278	0.858	8.645	9.833	0.385
3	0.182	0.070	0.160	0.394	0.279	0.303	10.152	8.349	0.225

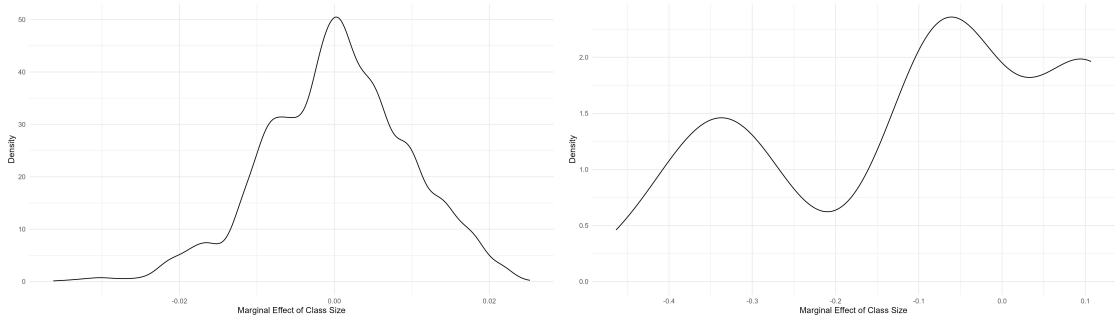
The table shows the means of observable teacher characteristics by group and treatment assignment and p-values testing the hypothesis of equality within each group.

Table C-9: Effect of Kindergarten Class Size on Later Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	Take Exam		Graduate HS		HS GPA	
Class size × Group 1	0.002	0.001	−0.002	−0.004	−0.081	−0.052
	(0.003)	(0.003)	(0.003)	(0.003)	(0.080)	(0.069)
Class size × Group 2	−0.013***	−0.012**	−0.010*	−0.006	−0.126	−0.097
	(0.005)	(0.005)	(0.005)	(0.005)	(0.103)	(0.106)
Class size × Group 3	0.003	0.006**	0.004	0.007*	0.035	0.033
	(0.003)	(0.003)	(0.003)	(0.004)	(0.084)	(0.076)
Nonwhite	0.011	0.031	−0.038	0.061	−1.013	-2.121***
	(0.022)	(0.038)	(0.029)	(0.046)	(0.632)	(0.837)
Free lunch	−0.295***	−0.289***	−0.187***	−0.186***	−3.261***	-3.647***
	(0.019)	(0.023)	(0.026)	(0.029)	(0.568)	(0.567)
Female	0.174***	0.174***	0.072***	0.076***	2.827***	2.901***
	(0.015)	(0.015)	(0.016)	(0.017)	(0.425)	(0.416)
Fixed Effects	Group	School	Group	School	Group	School
N	3813	3813	1878	1878	1485	1485

This table contains estimates from the same regressions as Table 7 in the main text and additionally displays the estimates for the demographic controls. Each column in the table displays the coefficients of OLS regressions on Kindergarten class size and controls for race, gender, and free lunch status. The coefficients on class size are allowed to vary by group membership. In columns (1)-(2), the dependent variable is a binary indicator for whether the student took a college entrance exam (SAT or ACT). In columns (3)-(4), the dependent variable is a binary indicator denoting graduation from high school. And in columns (5)-(6), the dependent variable is high school GPA. All standard errors are clustered at the level of the school. *p<0.1; **p<0.05; ***p<0.01 *p<0.1; **p<0.05; ***p<0.01

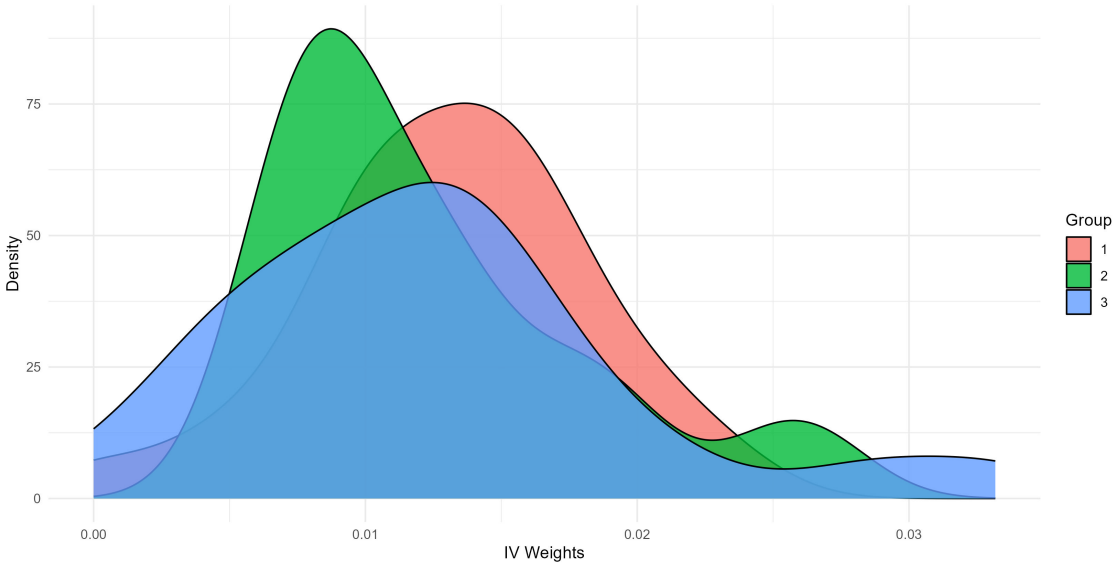
Figure C-1: Distribution of Estimated Posterior Marginal Effects



(a) Monte Carlo (b) STAR Data

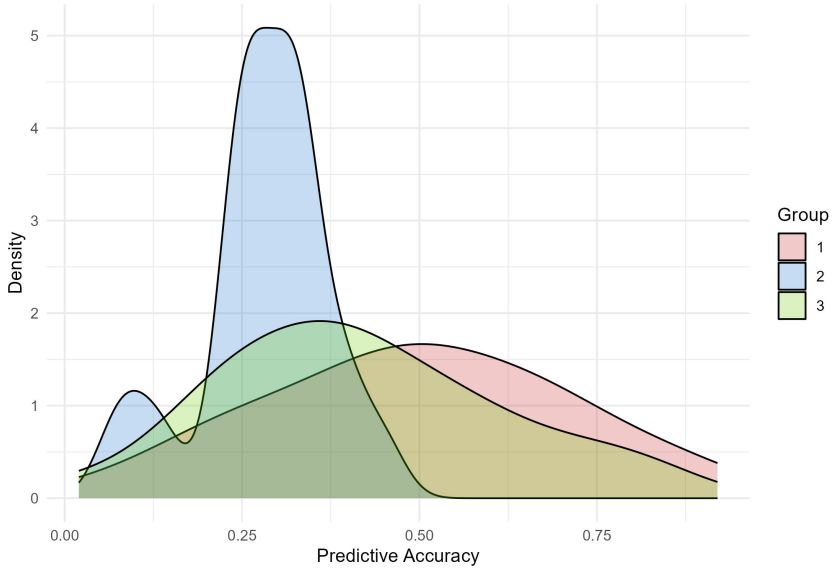
Figure (a) plots the density of school-specific posterior mean effects of class size on test scores from the Monte Carlo exercise described in section 6. The data is generated according to equation (13), which includes teacher and school random effects and effects of class size on test scores that are heterogeneous around a mean of zero. The EAMP algorithm is then applied to estimate the model in section 3 with $K = 2$ groups. The algorithm is therefore applied to a misspecified model. This figures plots the density of group-specific means obtained from estimation across 95 simulations. For comparison, Figure (b) plots the density of the school-specific posterior mean effects of class size on test scores across the 79 schools that participated in STAR according to the model with $K = 3$ groups.

Figure C-2: Density of School-Specific 2SLS Weights



The figure plots the density of weights across schools implied by the 2SLS estimand in equation (3). Separate densities are estimated for each group as uncovered the EAMP algorithm.

Figure C-3: Predictive Accuracy, Logit Models



The figure plots the distribution of predictive accuracy across all 79 schools in the STAR experiment. Each school's group membership is predicted according to a multinomial logit model with probabilities determined by Model 2 of Table 10 estimated on the remaining 78 schools. This figure plots the distribution of the model's predictive accuracy across all 79 schools. Predictive accuracy is defined as the fraction of 100 simulations in which the school is assigned to the correct group.

D Estimation Results From Sample Including Regular + Aide Classrooms

Table D-1: Class Size Marginal Effects by Group

Group	μ_k	Σ_k	$\sigma^2_{\epsilon,k}$	Schools	Students
1	-0.094	0.019	5.32	28	2046
2	-0.313	0.192	8.364	21	1609
3	0.056	0.001	10.652	30	2247
Avg. Effect	-0.097	0.081	8.180		
N	5902				

The table shows the estimated model parameters governing the effect of class size on test scores including controls for gender, race, and free lunch status. Controls are constrained to be equal across groups. The class size effects represent the effect of a one-unit increase in class size on test score performance.

Table D-2: Treatment Class Size Support by Group

Group	Class Size (Number of Students)					
	12	13	14	15	16	17
1	0.05	0.15	0.08	0.20	0.26	0.26
2	0.03	0.24	0.17	0.15	0.27	0.14
3	0.09	0.11	0.24	0.17	0.22	0.16

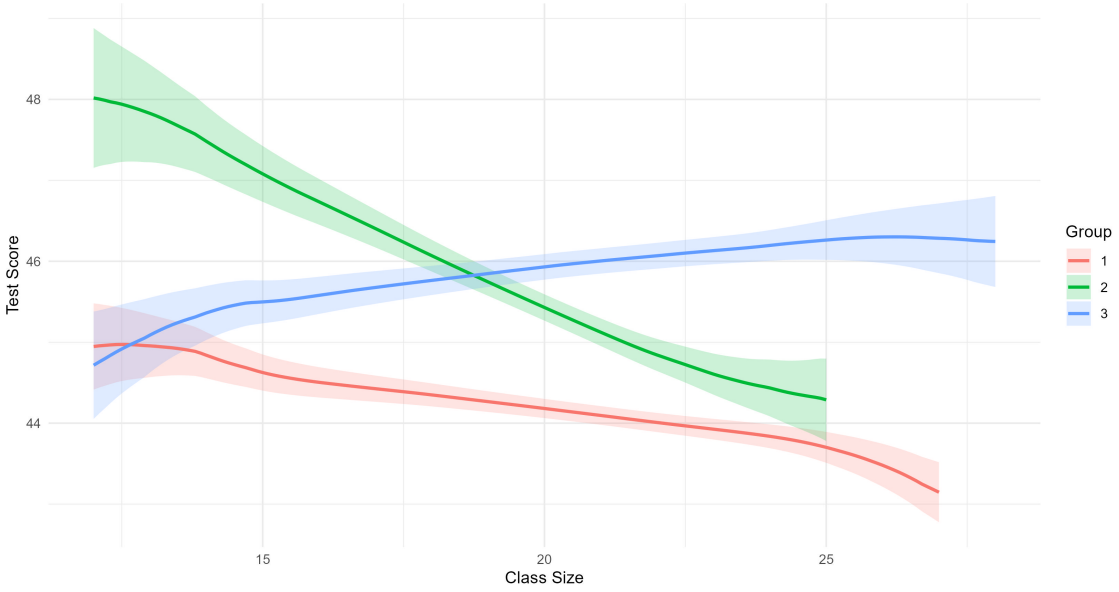
The table shows the Dirichlet prior means for each treatment class size by group. All parameters have been normalized by the sum of the estimated parameters for each group. Given this normalization, each number in the table can be interpreted as a group-specific prior on the fraction of observations in each cell. Zeros indicate that a particular group does not generate classes of that size.

Table D-3: Control Class Size Support by Group

Group	Class Size (Number of Students)													
	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	0.02	0	0.02	0.03	0.05	0.05	0.09	0.17	0.19	0.23	0.08	0.06	0.02	0
2	0	0	0.07	0.02	0.12	0.05	0.31	0.18	0.13	0.07	0.05	0	0	0
3	0	0.01	0	0	0.06	0.06	0.04	0.26	0.23	0.20	0.04	0.03	0.07	0.01

The table shows the Dirichlet prior means for each control class size by group. All parameters have been normalized by the sum of the estimated parameters for each group. Given this normalization, each number in the table can be interpreted as a group-specific prior on the fraction of observations in each cell. Zeros indicate that a particular group does not generate classes of that size.

Figure D-1: Nonparametric Regressions of Test Score on Class Size



The figure plots the semiparametric relationship between class size and test scores by group, as uncovered by the EAMP algorithm estimated on a sample that combines regular and regular + aide classrooms in the control group. All regressions use an Epanechnikov kernel and a bandwidth of 7.5. The regressions control for gender, race, and free lunch status.

E Estimation Results from Additional Specifications

Table E-1: Class Size Marginal Effects with K = 4 Groups

Group	μ_k	Σ_k	$\sigma^2_{\epsilon,k}$	Schools	Students
1	-0.090	0.028	4.642	27	1179
	(0.073)	(0.162)	(3.116)		
2	0.132	0.001	11.249	14	1132
	(0.085)	(0.096)	(2.585)		
3	-0.349	0.248	6.425	20	693
	(0.07)	(0.121)	(1.844)		
4	0.032	0.001	10.996	18	809
	(0.063)	(0.178)	(3.45)		
Avg. Effect	-0.094	0.105	8.276		
	(0.032)	(0.051)	(0.676)		
N	3813				

The table shows the estimated parameters governing the effect of class size on test scores for the model with $K = 4$ groups including controls for gender, race, and free lunch status. Controls are constrained to be equal across groups. The class size effects represent the effect of a one-unit increase in class size on test score performance. Bootstrapped standard errors from 95 bootstrapped data sets are in parentheses. Bootstrapping involves first sampling schools from the set of 79 schools that participated in the experiment and then sampling individuals within each school.

Table E-2: A Model with Group-Specific Coefficients on all Variables

	Group 1	Group 2	Group 3	Avg. Effect
μ_k	-0.086	-0.323	0.119	-0.102
Σ_k	0.020	0.147	0.00	0.089
$\sigma^2_{\epsilon,k}$	4.921	8.193	11.131	7.979
Female	0.385	0.614	0.917	0.630
Nonwhite	-1.285	-0.615	-1.237	-1.043
Free lunch	-1.286	-2.198	-1.717	-1.73
Schools	29	25	25	
Students	1325	1290	1198	
N		3813		

The table shows the estimated parameters for a model with $K = 3$ groups that allows the coefficients on the female, nonwhite, and free lunch binary variables to vary across groups.

Table E-3: Class Size Model Estimates, Listening Score

Group	μ_k	Σ_k	$\sigma^2_{\epsilon,k}$	Schools	Students
1	-0.023	0.003	5.905	33	1538
	(0.049)	(0.026)	(1.51)		
2	-0.306	0.063	7.479	19	917
	(0.076)	(0.074)	(1.57)		
3	0.112	0.002	10.193	27	1320
	(0.075)	(0.048)	(2.07)		
Avg. Effect	-0.045	0.043	7.787		
	(0.029)	(0.029)	(0.50)		
N	3775				
Dependent Variable Mean & SD	53.83	3.33			

The table shows the estimated model parameters governing the effect of class size on listening test scores. The table is analogous to table 2 in the main text. The model controls for gender, race, and free lunch status, and their coefficients are constrained to be equal across groups. The class size effects represent the effect of a one-unit increase in class size on the Stanford listening score. Bootstrapped standard errors from 95 bootstrapped data sets are in parentheses. Bootstrapping involves first sampling schools from the set of 79 schools that participated in the experiment and then sampling individuals within each school.

Table E-4: Group Comparison, Main Estimates and Listening

		Listening Score Group		
		1	2	3
Main Group	1	22	3	6
	2	5	13	5
	3	6	3	16

The table compares the clustering of schools into groups in the main model with the model in which listening scores serve as the dependent variable.

Table E-5: Class Size Model Estimates, Self Concept Score

Group	μ_k	Σ_k	$\sigma^2_{\epsilon,k}$	Schools	Students
1	0.086	0.003	42.177	20	909
	(0.083)	(0.181)	(16.22)		
2	-0.457	0.145	17.498	27	1239
	(0.111)	(0.161)	(3.19)		
3	0.096	0.352	14.189	26	1140
	(0.080)	(0.141)	(20.15)		
Avg. Effect	-0.115	0.248	23.174		
	(0.053)	(0.095)	(2.02)		
N	3288				
Dependent Variable Mean & SD	56.03	5.30			

The table shows the estimated model parameters governing the effect of class size on self concept scores. The table is analogous to table 2 in the main text. The model controls for gender, race, and free lunch status, and their coefficients are constrained to be equal across groups. The class size effects represent the effect of a one-unit increase in class size on the SCAMIN self concept score. Bootstrapped standard errors from 95 bootstrapped data sets are in parentheses. Bootstrapping involves first sampling schools from the set of 73 schools with self-concept scores that participated in the experiment and then sampling individuals within each school. Due to the higher prevalence of missing self concept scores than for cognitive skill measures, this analysis only featured 73 schools.

Table E-6: Group Comparison, Main Estimates and Self Concept Scores

		Self Concept Score Group		
		1	2	3
Main Group	1	9	6	11
	2	5	9	8
	3	6	12	7

The table compares the clustering of schools into groups in the main model with the model in which self concept scores serve as the dependent variable.