A Reappraisal of the Effect of Class Size on Children's Learning*

Karun Adusumilli Francesco Agostinelli

Emilio Borghesan

November 2022

Abstract

The Student-Teacher Achievement Ratio (STAR) Project is one of the most important educational experiments in the United States, and its results have been used to promote schooling reforms. In this paper, we argue that the simple experimental evaluation of STAR has limited policy relevance due to its experimental design. While children were randomized into classes of different sizes, the intensity of class size reduction was subject to schools' heterogeneous compliance responses, which can confound treatment effects with selection on gains. In light of this experimental feature, we re-evaluate the class size effects in STAR via a new econometric framework that allows for heterogeneous treatment effects and possible grouped selection on targeted treatment and control class sizes. We find that a small group of schools displays large benefits from class size reduction, while the vast majority did not. Our results suggest that policy proposals to reduce the pupil/teacher ratios should be implemented in small scale and in a targeted modality.

JEL Classification: C51, I2, H52, J24

*We thank Michael Dinerstein for useful comments and suggestions. 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 Pennsylvania, 133 S 36th St, Philadelphia, PA 19104, USA (email: borghesa@sas.upenn.edu).

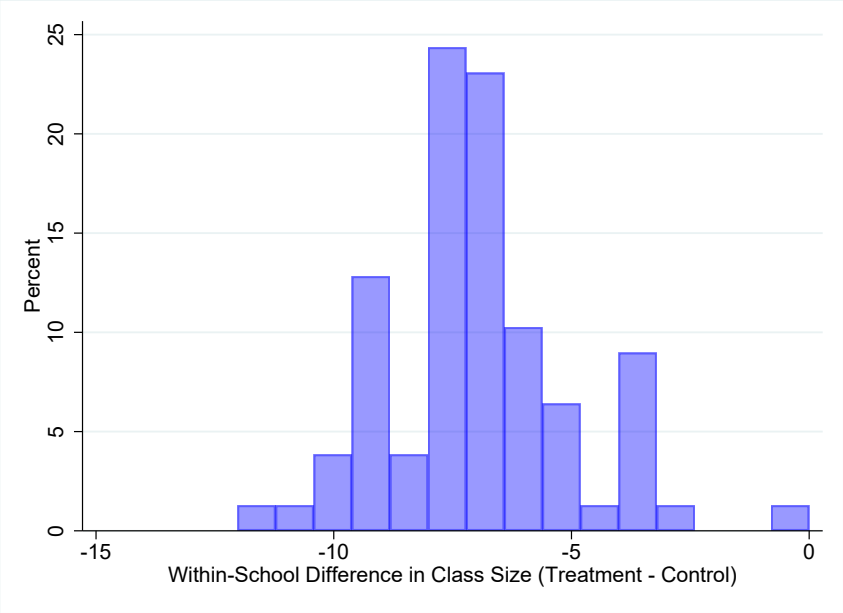
1 Introduction

Recent empirical work on the effect of class size on student learning has animated the US policy debate on public educational investments and the extent to which public schools should increase teacher hiring. While early observational studies pointed towards a small or null effect of class size on academic performance, the preliminary evaluations of Tennessee's Student Teacher Achievement Ratio (STAR) experiment led many researchers to conclude that reducing class size generated causal gains in student learning (Folger and Breda 1989, Finn and Achilles 1990, Word et al. 1990, Schanzenbach 2006). The STAR experiment, conducted between 1985 and 1989, randomized kindergarten students into one of three different class types: small, regular, and regular with a teacher's aide. Researchers have found that students in the small class size experienced significant gains in test scores after a single year relative to students in the other two class types. Because of its experimental design, some have argued that the STAR experiment provides the only unbiased empirical evidence of class size effects on children, and for this reason the results were used to advocate for universal policies for hiring more public school teachers.[1]

In this paper we provide a reappraisal of the policy relevance of the STAR class size effects. In doing so, we present three main contributions. We show that the STAR experimental design allowed for heterogeneity in schools' compliance behaviors, with some schools generating substantially larger differences in class size between treatment and control groups than other schools. In such a setting, converting the reduced-form experimental estimates into the policy-relevant effects of class size on test scores is not straightforward unless there is no heterogeneity in the marginal returns to class size across schools. Our analysis, however, provides evidence of significant effect heterogeneity across schools. Second, we develop and apply a new econometric method that accounts for heterogeneous returns of class size, as well as possible *grouped* selection on unobserved gains. Finally, our empirical analysis suggests that the STAR experimental effects are largely driven by one specific group of schools—which represents approxi-

[1]The academic debate lead to the publication of a book titled "The class size debate" (Mishel and Rothstein 2002) where professors Alan Krueger and Eric Hanushek argue on the class size effects on student achievements.

Figure 1: Heterogeneity in Class-Size Reduction Compliance



The figure plots the distribution of class size reduction between regular and small classes for different schools in the Tennessee STAR experiment.

mately thirty percent of the schools in the sample—a result that highlights the importance of targeting educational policies instead of adopting a universal approach.

The first contribution of this paper is to highlight how the institutional details of the STAR experimental design effectively limit the interpretation of the reduced-form experimental results as policy-relevant treatment parameters for class size reductions. Figure 1 shows the heterogeneity in class size compliance behavior among schools. While on average class size is reduced by approximately seven students in the experiment, the reduction is heterogeneous across schools, ranging between two and thirteen students. This heterogeneity in compliance behavior is a result of the design of the STAR experiment, which randomized students into one of three class types within the same school (small, regular, and regular with a teacher's aide) and let schools freely set the target size of a particular class type in a nonrandom fashion. Schools' compliance behavior may potentially be driven by pressure from parents or decisions made by school principals, which could generate selection on the unobserved gains from reducing class size. Con-

sistent with this theory, we find that schools whose students benefit the most from reducing class size had the smallest difference in class size between treatment and control classrooms. Our results show that, although the within-school experimental evaluation remains internally valid, using the experimental randomization as an instrumental variable to identify the policy-relevant marginal effect of class size on test scores (as in Krueger 1999) does not in fact identify the targeted parameter of interest, unless the class size effects are effectively homogeneous across schools.

Our second contribution is to develop a new econometric framework that allows for heterogeneous class size effects among students together with possible selection of class sizes based on unobserved gains. We allow for the distributions of class sizes and treatment effects to vary by school, and we posit that a low-dimensional set of latent parameters govern their joint distribution across schools. Our approach effectively classifies schools into groups based on similar values of these latent parameters, so that within each group, the distribution of class sizes and the marginal effects of class size on test scores are independent, rendering the assignment of class sizes plausibly random. However, these distributions can vary across groups, thereby allowing for grouped selection on unobserved gains. We use a Grouped Random Effects (Adusumilli 2020) methodology to simultaneously classify schools into groups and compute parameters governing the distribution of class sizes and marginal treatment effects. The number of groups is determined through an information criterion (AIC).

Finally, our third contribution relates to the policy implications of our results. We find that all of the test score gains for small classes in the STAR experiment are driven by a single group of schools that comprise 30% of the sample. We conduct counterfactual simulations that show how a targeted and smaller-scale version of the same intervention with one-third of the original budget would have generated the same average effects as the STAR experiment. The schools that benefit the most from class size reductions tend to be heavily segregated by race and contain high rates of student poverty. We also show that the STAR project reduced the Black-white test score gap, and that the same targeted intervention would have achieved 72.5% of this Black-white gap reduction. The vast heterogeneity

in the effects of class size on academic performance is consistent with the previous contrasting conclusions from the literature: While the majority of children may have little if any benefit from class size reduction, children in socioeconomically disadvantaged schools tend to benefit greatly from reducing pupil-teacher ratios. Overall, our results suggest that policies that aim to increase the hiring of new teachers at public schools in the service of reducing class size should be targeted to the schools that benefit the most from it.

Related Literature. An extensive literature uses observational and quasi-experimental designs to identify the impact 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 ratio).[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 the Israeli public schools and find that class size reductions show positive effects on fourth and fifth graders, but not on third graders. Hoxby (2000) exploits longitudinal idiosyncratic variation in the student population to identify the effects of class size on student achievement. She finds no evidence of any positive effect of class reduction on children learning. Rivkin, Hanushek, and Kain (2005) compare the relative importance of teachers and school quality on children's achievement. They conclude that improving teacher quality is more effective and efficient than reducing class size in improving students' achievements.

An important push-back to this conclusion arose from subsequent analysis of the STAR experiment. Multiple evaluations of this experiment find positive shorter-term and longer-term estimates for children who were randomly assigned to a smaller class. Because of the experimental nature of these studies, the results

[2] Card and Krueger (1992b) find that that 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.

have been used to reinterpret previous research on the topic, and to provide strong arguments in support of policy proposals to reduce the pupil/teacher ratios in schools via public-school teacher hiring (see Krueger 1999, Krueger and Whitmore 2001, Schanzenbach 2006, Chetty et al. 2011).

In this paper we highlight that the desirability of the research design alone—experimental assignment of class types—does not guarantee the identification of the policy-relevant class size effects on academic performance. The majority of studies based on the STAR experiment identify the reduced-form effects of the experimental randomization on various outcomes. While these reduced-form evaluations are informative about the causal effects of the randomization in the STAR experiment, they are generally silent about the effects of class size reductions on test scores. This is due to heterogeneity in the behavioral compliance in class size reduction between treatment arms (see Figure 1). For example, Krueger 1999 try to estimate the class size effect for children in STAR using the original experimental assignment as an instrumental variable for the actual class size. However, we argue that the heterogeneity in the endogenous class size compliance behavior of schools makes it difficult to give a causal interpretation to the IV estimates in STAR due to possible selection on unobserved gains. Our study is the first to our knowledge that highlights how the limitations in the STAR experiment make the comparison between the STAR estimates and the class size effects of the previous literature not clear-cut. Moreover, we provide a new econometric framework that allows us to identify the policy relevant class size effects by accounting for heterogeneous returns of class size in the population, as well as possible selection on unobserved gains across schools.[3]

Our work also relates to the growing literature that studies the ability of field experiments to inform policy decisions at scale. Jepsen and Rivkin (2009) evaluate the impact of California's billion-dollar class size-size-reduction program on student achievement, a program that induced the state of California to search and hire a large number of new teachers to implement the targeted class size reduction. They find that the size of the program caused the state to hire teachers with neither prior experience nor full certification, and this "dampened the benefits of

[3] Krueger and Whitmore (2001) used the experimental variation as an instrument to estimate the effect of class size on the probability of taking the ACT/SAT.

smaller classes, particularly in schools with high shares of economically disadvantaged, minority students." Al-Ubaydli, List, and Suskind (2020) define this threat to scalability as the diseconomies of scale in participation and compliance costs. Because our results suggest that a small fraction of schools drive the vast majority of the STAR treatment effects, a targeted class-size intervention would also represent a more scalable program as hiring a smaller number of teachers helps to preserve their quality.

Finally, 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 (2017) 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. In the present context, unobserved heterogeneity corresponds to parameters governing the target sizes for treatment and control classrooms and the experimental effects of the randomization on test scores. In this paper, we use Grouped Random Effects (GRE), where the distribution of unobserved heterogeneity is allowed to vary across groups. Clearly, GFE is a special case of this. We use GRE over GFE for three reasons: First, it enables us to jointly model the treatment and control class sizes together with the treatment effects. By modeling these jointly, we are able to account for the heterogeneity in behavioural responses, and also their correlation with the treatment effect. The typical GFE model does not model class sizes. Second, we are interested in various measures relating to the distribution of unobserved heterogeneity. In this regard, GRE estimates are more precise as they require fewer groups. In GFE, observations are clustered so that the unobserved heterogeneity is approximately constant within each group. By contrast, GRE forms groups so that the unobserved heterogeneity is approximately independent of the covariates within each group. This is a weaker requirement, and it therefore allows for fewer groups (or, equivalently, smaller bias with the same number of groups). Furthermore, we can also use posterior averaging to get better estimates of the marginal effects. Third, the GRE method simultaneously computes both the group assignments and estimates of group specific parameters. By contrast, Bonhomme, Lamadon, and Manresa (2017) suggest a two step method, which requires specifying moments for the first stage group assign-

ments, and can be less accurate than one step methods. Bonhomme and Manresa (2015) suggest a one step method, but this is only proposed for linear panel data models (our setting is a nonlinear one due to the need to model class sizes).

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).

2 The STAR Experiment and a Simple Motivational Framework

2.1 The Aim of This Paper

The goal of this paper is to understand the relationship between class size and student test scores. We specify the following linear model for the test score of student i in school s,

$$y_{is} = \eta_s + \alpha_{is} n_{is} + x_{is}'\theta + \varepsilon_{is} \ , \tag{1}$$

where y_{is} denotes the test score, η_s is a constant that varies by school, n_{is} denotes the number of students in student i's classroom, x_{is} is a vector of observed controls, and ε_{is} is a random shock. The school fixed effects are included in the model because of the school-level randomization of students into small and regular-sized classrooms. We define the average marginal effect of class size on student performance as one of the main parameters of interest ($E[\alpha_{is}]$).

Our goal is to use the Tennessee STAR experiment to obtain an estimate of this policy-relevant parameter and to learn something about the heterogeneity of α_{is} across schools. Prior analysis of the experiment has shown that students randomized into smaller classes perform significantly better than their peers in large classes. However, recovering an estimate of the marginal returns to class size from the experiment is considerably more challenging.

The experimental design does not ensure that $\alpha_{is} \perp\!\!\!\perp n_{is}$. Instead the STAR experiment randomized students into small and regular-sized classrooms, leaving principals at each school free to determine a target size for their treatment

and control classes. Various factors – including the number of classrooms in the school and the size of the cohort – may have resulted in different target sizes for the treatment and control classes. In some schools, these target sizes may have been determined so as to minimize the tension that could result from a large perceived difference in resources targeted towards students in different classrooms.[4] Such considerations would violate the usual assumption of no-selection on unobserved gains $\alpha_{is} \perp\!\!\!\perp n_{is}$. In the next section we discuss the possible consequences for the causal interpretation of the Instrumental Variables (IV) estimand in this context.

2.2 The Limitations of Instrumental Variables in Our Setting

Prior studies (Krueger 1999) have used randomization into a treatment classroom as an instrument for class size within an IV framework. In this section, we provide a simple example to show how IV fails to recover a policy-relevant treatment parameter for class size.

Consider the case of a set of schools $s \in \mathcal{S} := \{1, 2, \ldots, S\}$ in the STAR experiment, each having one small class ($c = 1$) and one large class ($c = 2$). Students are randomized into each of the class types, and each class in school s is characterized by its own size $n_{c,s}$, where $n_{1,s} < n_{2,s}$.

As in Krueger (1999), we use the original within-school class-type randomization ($Z_{is} \in \{0, 1\}$) as an instrument for class size, n_{is}, in the following simplified model:

$$y_{is} = \eta_s + \alpha_{is} n_{is} + \epsilon_{is} \ . \tag{2}$$

Relative to equation 1, equation 2 omits the controls, x_{is}. We rewrite equation 2 after de-meaning y_{is} and n_{is} of their school-specific means (within-school trans-

[4] Krueger (1999) documents that tension between parents and school principals led to both attrition from control classes and re-randomization in later years of the experiment. Many students in initially large classes were moved to small classes, and vice versa, to appease parents. It is not inconceivable that such concerns would affect the determination of treatment and control classes at the experiment's inception.

formation),

$$\Delta y_{is} = \alpha_{is}\Delta n_{is} + (\epsilon_{is} - \bar{\epsilon}_{is}), \tag{3}$$

where $\Delta n_{is} = (1-\phi_{1,s}) \cdot (n_{1,s} - n_{2,s}) \times \mathbb{1}\{c_i = 1\} + \phi_{1,s} \cdot (n_{2,s} - n_{1,s}) \times \mathbb{1}\{c_i = 2\}$, $\phi_{1,s} = \frac{n_{1,s}}{n_{1,s}+n_{2,s}}$, and c_i denotes the classroom assignment of individual i. In this case, the transformed instrumental variable is

$$\Delta Z_{i,s} \equiv (Z_{i,s} - \bar{Z}_s) = \begin{cases} -\phi_{1,s} & \text{if } Z_{i,s} = 0 \\ (1-\phi_{1,s}) & \text{if } Z_{i,s} = 1 \end{cases}, \tag{4}$$

where schools have different shares, $\phi_{1,s}$, of students in the small class depending on their compliance behavior $(n_{1,s} - n_{2,s})$ and their school size $(n_{1,s} + n_{2,s})$.

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: $E[\epsilon_{is}|Z_{is}, s] = E[\epsilon_{is}|s]$.*

In this framework, even if our instrument is relevant and randomly assigned, we show that the IV estimand does not identify the policy-relevant treatment effect of interest, but instead it identifies a weighted average effect, where the weights are a function of the possible selection on unobserved gains of schools in reducing class size.

Proposition 1. *Suppose Assumption 1 holds and that the original STAR randomization is relevant, i.e., it generates significant differences in class size within each school. The IV estimand of model (2) identifies a weighted average effect of class size reduction, with weights that depend on the endogenous compliance behavior of schools with respect to the class size reduction in the experimental setting*

$$\alpha_1^{IV} = \frac{E(\Delta y \Delta Z)}{E(\Delta n \Delta Z)} \equiv \sum_{s \in \mathcal{S}} \alpha_s \frac{\phi_s \phi_{1,s}(1-\phi_{1,s})(n_{1,s} - n_{2,s})}{\sum_{s \in \mathcal{S}} \phi_s \phi_{1,s}(1-\phi_{1,s})(n_{1,s} - n_{2,s})}, \tag{5}$$

where ϕ_s represents the share of children attending school s and $\alpha_s = \mathbb{E}[\alpha_{is} \mid s]$.

Proposition 1 reveals that the IV estimand does not generally identify the policy relevant treatment parameter, with the exception of two special cases: (i) when treatment effects are homogeneous in the populations $\alpha_s = \alpha$ for all $s \in \mathcal{S}$; or (ii), when the compliance behavior of schools is independent of the return to class size reduction $(\phi_{1,s}, \Delta n_{is}) \perp\!\!\!\perp \alpha_s$. However, both assumptions impose strong restrictions on either the treatment effect heterogeneity or on the schools' objective functions. More generally, the IV estimand does not admit a causal interpretation—even if researchers can observe the counterfactual classroom outcome for each school in the experiment—because schools may 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.

Moreover, with a fixed size for each cohort, an increase in the size of one class necessitates a reduction in the size of the other class. This dependence causes the IV weights in 5 to not be monotone in the intensity of the class size reduction, $n_{1,s} - n_{2,s}$.

3 From STAR Randomization to Policy-Relevant Effects

In this section, we outline an approach that can simultaneously overcome the shortcomings of IV and enable us to learn about the heterogeneity in α_{is}. Our approach is robust to both nonindependence of α_{is} and n_{is} and functional form misspecification in 1. It will cluster schools together into a fixed number of groups, $k = 1, \ldots, K$, based on a common set of underlying parameters that jointly determine the distribution of class sizes within each school and the treatment effects generated by the experiment. After performing this grouping, we will be able to estimate a version of 1 that yields an unbiased estimate of $\mathbb{E}[\alpha_{is}]$ and reveals the extent of heterogeneity in $\mathbb{E}[\alpha_{is} \mid s]$ across schools.

3.1 Reduced Form Test Score Equation

The test scores generated by the STAR experiment are determined by the following reduced-form equation for student i in school s belonging to group k:

$$y_{isk} = \eta_s + \beta_{isk} Treat_{isk} + x_{isk}{}'\theta + \epsilon_{isk} \ , \tag{6}$$

$$\beta_{isk} \sim N(\mu_k, \Sigma_k) \ , \tag{7}$$

where $Treat_{isk}$ denotes whether individual i is randomized into a small (treated) classroom and x_{isk} denote observable student characteristics. In the model, treatment effects are heterogeneous and are assumed to follow a Normal distribution with a mean, μ_k, and variance, Σ_k, that are allowed to differ by group.[5] 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 constant variance, σ^2_ϵ, for all students.

Instead of the above specification, one could imagine using the class size, n_{is}, in place of $Treat_{isk}$ and directly estimating the effect of this on test scores. However, if class size has a nonlinear impact on test scores, imposing a linear specification may imply that the clustering algorithm sorts schools on the basis of non-linear effects rather than on the basis of heterogeneity in the reduced form effect of the treatment and the chosen class size difference. Hence this alternative specification is less robust. By contrast, our specification can be interpreted as the intention to treat effect, if we think of $Treat_{isk}$ as an instrumental variable and n_{is} as the treatment. We assume that there is within-group independence between the observed class size difference and the reduced form impact of treatment (but these quantities may be correlated across groups).

In practice we are interested in the effect of reducing the class size by 1 unit on test scores. We obtain this effect by running regressions of y_{isk} on n_{isk} for all the schools within a cluster, following the implementation of our clustering algorithm.

[5] Even if interest is on the mean effects rather than the variances, the iterative algorithm we describe in the next section will successively update μ_k based on estimates of Σ_k, so it is necessary to estimate both the mean and the variance.

3.2 Class Size Model

Each school chooses a vector of class sizes for the treated and control groups denoted by $\boldsymbol{n}_{sk}^{(t)}$ and $\boldsymbol{n}_{sk}^{(c)}$.[6] We posit that the following multinomial model generates these class sizes:

$$\begin{aligned} \boldsymbol{n}_{sk}^{(c)} &\sim \text{multinomial}(p_{sk}^{(c)}) \text{ for all } k \text{ in the control group,} \\ \boldsymbol{n}_{sk}^{(t)} &\sim \text{multinomial}(p_{sk}^{(t)}) \text{ for all } k \text{ in the treatment group.} \end{aligned} \tag{8}$$

Here, $p_{sk}^{(c)} \equiv \{p_{sk1}^{(c)}, \ldots, p_{skL}^{(c)}\}$, $p_{sk}^{(t)} \equiv \{p_{sk1}^{(c)}, \ldots, p_{skM}^{(c)}\}$ 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, $L = 6$ 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 (treatment/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. The assumption that $p_{sk}^{(c)}$ and $p_{sk}^{(t)}$ are school-specific rather than classroom-specific is consistent with the random assignment of students and teachers to classrooms.

Importantly, we model not just the difference in class sizes (between the treatment and control groups) but also the levels of the class sizes as well. It is quite plausible that schools with different control class sizes are associated with different treatment effects e.g., schools with large classrooms to begin with may respond more strongly to a reduction in class size. At the same time, the original size of the classes may also directly influence the class size difference chosen in the experiment, thereby inducing varying degrees of compliance that could be based on selection on gains (e.g., schools which expect a large treatment effect

[6] Note that class sizes in the individual test scores model, n_{isk}, are a scalar integer-valued variable, while the class size vectors, which lack an i subscript, are vectors of counts denoting the number of classes in school s of each size.

may elect for a larger class size difference). By taking the levels into account, we are able to jointly model the correlation between the initial class size, the chosen class size difference and the reduced form impact of treatment.

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

3.3 Endogeneity and Grouped Random Effects

Endogeneity may arise in equation 1 if a common set of parameters determine both the returns to class size, α_{is}, and the class sizes themselves, n_{is}. One can imagine that there exist unobserved school-specific latent variables, ξ_s, simultaneously affecting both α_{is} and $(p_{sk}^{(c)}, p_{sk}^{(t)})$. Instead, β_{isk} in equation 7 is not endogenous because of the experimental design. However, the treatment effect β_{isk} represents both the marginal effects of class size on test scores, α_{is}, as well as the within-school deviations in class sizes generated by the experiment.

We adopt a Grouped Random Effects (GRE) approach, following Adusumilli (2020), to group schools according to a common set of parameters, ξ_s, that jointly determine class sizes, through $(p_{sk}^{(c)}, p_{sk}^{(t)})$, and the treatment effects, β_{isk}. We assume that ξ_s is constant within each group, but even if ξ_s is continuous, the GRE estimator allows us to discretize its support, ensuring that ξ_s is at least approximately constant in each group. This approach specifies and estimates priors over the random parameters for each group, $(\beta_{isk}, p_{sk}^{(c)}, p_{sk}^{(t)})$. By recovering parameters governing the distribution of $(p_{sk}^{(c)}, p_{sk}^{(t)})$, we are able to isolate the intensity of the class size reduction between treatment and control classes within each school. The variation in β_{isk} that is not determined by variation in class size between treatment and control classes is generated by variation in the marginal returns to class size, α_{is}.

We use GRE as it enables us to jointly model both treatment effects and class sizes. By contrast, a Grouped Fixed Effects model as in Bonhomme and Manresa (2015) would only try to group schools based on the differences in β_{isk}, and fail to

capture the correlation between $\beta_{isk}, \boldsymbol{n}_{sk}^{(c)}$ and $\boldsymbol{n}_{sk}^{(t)}$. Hence it cannot fully account for the heterogeneity in compliance and the corresponding selection on gains.

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

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) := \prod_s \prod_{i:s(i)=s} \prod_k N(\beta_{isk}|\mu_k, \Sigma_k)^{w_s(k)}, \tag{9}$$

so that the student-specific treatment effects, β_{isk}, are random draws from a group-specific normal distribution with mean μ_k and variance Σ_k.[7] This assumes that the distribution of student and school characteristics affecting test scores is similar across all schools within the same group.

For the multinomial probabilities $\boldsymbol{p}^{(c)} \equiv \{p_{sk}^{(c)}\}_{s=1}^S, \boldsymbol{p}^{(t)} \equiv \{p_{sk}^{(t)}\}_{s=1}^S$, we employ a group-specific Dirichlet prior

$$\begin{aligned} \pi(\boldsymbol{p}^{(c)}|\boldsymbol{\eta}^{(c)}) &= \prod_s \prod_k \text{Dirichlet}(p_{sk}^{(c)}|\eta_k^{(c)})^{w_s(k)}, \\ \pi(\boldsymbol{p}^{(t)}|\boldsymbol{\eta}^{(t)}) &= \prod_s \prod_k \text{Dirichlet}(p_{sk}^{(t)}|\eta_k^{(t)})^{w_s(k)}, \end{aligned} \tag{10}$$

where $\boldsymbol{\eta}^{(c)} \equiv \{\eta_k^{(c)}\}_{k=1}^K, \boldsymbol{\eta}^{(t)} \equiv \{\eta_k^{(t)}\}_{k=1}^K$ are group-specific parameters, and $\text{Dirichlet}(p|\eta)$ denotes the pdf of the Dirichlet distribution with parameter η evaluated at p. The Dirichlet distribution is the conjugate prior for the multinomial class size distributions $(p_{sk}^{(c)}, p_{sk}^{(t)})$. It is chosen both for computational tractability, and because it is a distribution over probability distributions on the unit simplex. It is, therefore, an appropriate choice for generating probabilities of proportions of control and treatment class sizes at each school.

3.4 Probability model

Let $\boldsymbol{y} \equiv \{y_{isk}\}_{i,s,k}$, $\boldsymbol{x} \equiv \{x_{isk}\}_{i,s,k}$, and $\boldsymbol{T} \equiv \{Treat_{isk}\}_{i,s,k}$ denote the vectors of test scores, covariates, and treatment indicators. Given the treatment effect hetero-

[7] While our baseline model has only one random coefficient, we write the model in a more general form allowing for multiple random coefficients and nondiagonal covariance matrices.

geneity $\boldsymbol{\beta}$, the treatment indicators, $\boldsymbol{T}$, and the class size vectors $\boldsymbol{n}^{(c)} := \{n_{isk}^{(c)}\}_{i,s,k}$, and $\boldsymbol{n}^{(t)} := \{n_{isk}^{(t)}\}_{i,s,k}$ we model test scores using the quasi-log-likelihood

$$\begin{aligned}\ln p(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{n}^{(c)}, \boldsymbol{n}^{(t)}, \boldsymbol{T}, \boldsymbol{x}, \theta) &:= \sum_{s} \sum_{i:s(i)=s} \ln p(y_{isk}|\beta_{isk}, Treat_{isk}, x_{isk}, \theta) \\ &:= \sum_{s} \sum_{i:s(i)=s} \{y_{isk} - \beta_{isk} Treat_{isk} - \theta x_{isk} - \eta_s\}^2 \ .\end{aligned}$$

Note that the above is a quasi-likelihood since it assumes ϵ_{isk} has a standard normal distribution, which we do not know to be necessarily true. However, as noted in Bonhomme and Manresa (2015), due to the linear structure of the model for test scores, estimating the quasi-likelihood still yields consistent estimates of the model parameters. The model is also correctly specified if we assume that $\text{Var}[\epsilon_{isk}]$ is independent of the group identity.

The multinomial class size model in (8) implies that the distribution of the sizes of treatment and control classrooms is given by

$$\ln p(\boldsymbol{n}^{(c)}|\boldsymbol{p}^{(c)}) := \sum_{s=1}^{S} \sum_{g=1}^{N^{(c)}(s)} \sum_{k=1}^{K} w_s(k) \left\{ \sum_{l=0}^{L} \mathbb{I}\{n_{skg}^{(c)} \equiv 12 + l\} p_{skl}^{(c)} \right\} , \tag{11}$$

$$\ln p(\boldsymbol{n}^{(t)}|\boldsymbol{p}^{(t)}) := \sum_{s=1}^{S} \sum_{g=1}^{N^{(t)}(s)} \sum_{k=1}^{K} w_s(k) \left\{ \sum_{m=0}^{M} \mathbb{I}\{n_{skg}^{(t)} \equiv 16 + m\} p_{skm}^{(t)} \right\} , \tag{12}$$

where $N^{(c)}(s)$ and $N^{(t)}(s)$ are the observed number of control and treatment classes, respectively, in school s. The expressions in (11) and (12) 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 times the group-specific multinomial probabilities for the observed class sizes. We do not model the assignment of treatment and control status to the classroom, since these were determined entirely randomly in the experiment.

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

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

Additionally, in view of (9) and (10), 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 prior parameters:

$$\begin{aligned}
&\max_{\{w_s(k)\},\{\boldsymbol{\rho}_k\}} \ln \int p(\boldsymbol{y}, \boldsymbol{n}^{(c)}, \boldsymbol{n}^{(t)} | \boldsymbol{\alpha}, \boldsymbol{T}, \boldsymbol{x}, \theta) \pi\left(\boldsymbol{\alpha}|\boldsymbol{\rho}\right) d\boldsymbol{\alpha} \\
&= \max_{\{w_s(k)\},\{\boldsymbol{\rho}_k\}} \sum_{s=1}^{S} \sum_{k=1}^{K} w_s(k) \ln \int p(\boldsymbol{y}_s, \boldsymbol{n}_s^{(c)}, \boldsymbol{n}_s^{(t)} | \boldsymbol{\alpha}_s, \boldsymbol{T}, \boldsymbol{x}, \theta) \pi\left(\boldsymbol{\alpha}_s|\boldsymbol{\rho}_k\right) d\boldsymbol{\alpha}_s. \qquad (13)
\end{aligned}$$

Following Adusumilli (2020), we solve the above maximization problem using the Expectation, Assignment, Maximization, and Propagation (EAMP) algorithm. To use the algorithm, we follow Neal and Hinton (1998) and rewrite the expected likelihood in (13) as follows:

$$\begin{aligned}
&\max_{\{w_s(k)\},\{\boldsymbol{\rho}_k\}} \sum_{s=1}^{S} \sum_{k=1}^{K} w_s(k) \ln \int p(\boldsymbol{y}_s, \boldsymbol{n}_s^{(c)}, \boldsymbol{n}_s^{(t)} | \boldsymbol{\alpha}_s, \boldsymbol{T}, \boldsymbol{x}, \theta) \pi\left(\boldsymbol{\alpha}_s|\boldsymbol{\rho}_k\right) d\boldsymbol{\alpha}_s \\
&= \max_{\substack{\{q_{sk}(\cdot)\}, \\ \{w_s(k)\},\{\boldsymbol{\rho}_k\}}} \sum_{s=1}^{S} \sum_{k=1}^{K} w_s(k) \left\{ E_{q_{sk}(\cdot)} \left[\ln p(\boldsymbol{y}_s, \boldsymbol{n}_s^{(c)}, \boldsymbol{n}_s^{(t)} | \boldsymbol{\alpha}_s, \boldsymbol{T}, \boldsymbol{x}, \theta) \right] - \mathrm{KL}\left(q_{sk}(\boldsymbol{\alpha}_s) \,||\, \pi(\boldsymbol{\alpha}_s|\boldsymbol{\rho}_k)\right) \right\},
\end{aligned} \qquad (14)$$

where $q_{sk}(\cdot)$ denotes a distribution over $\boldsymbol{\alpha}_s$ that is (potentially) group specific, 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

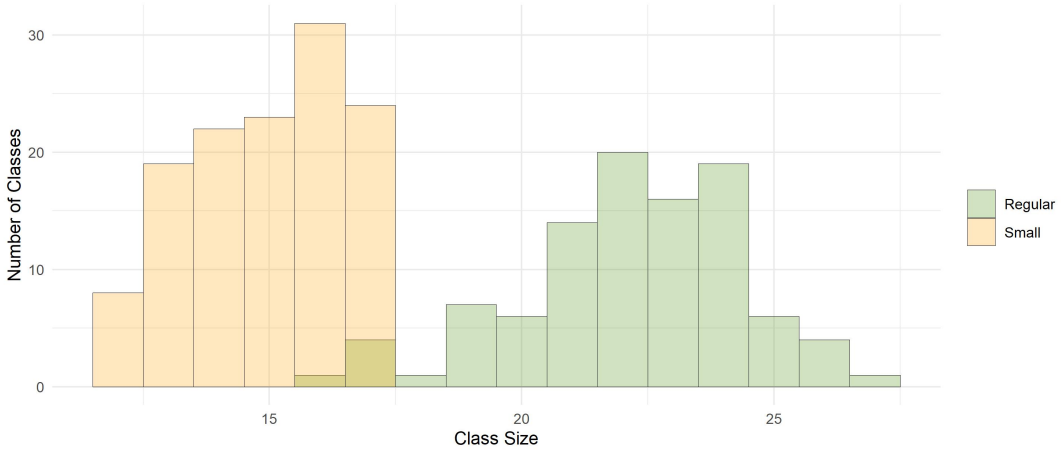
$\{q_{sk}(\cdot)\}, \{w_s(k)\}, \{\boldsymbol{\rho}_k\}$ holding other quantities fixed. This results in a sequence of four steps – Expectation, Assignment, Maximization, and Propagation – that are repeated in an iterative process until the algorithm converges. Note that, we first demean all variables by their school-specific means to eliminate the school fixed effects before running the estimation algorithm. A detailed description of the various steps of the algorithm is provided in the appendix.

5 Data

We conduct our analysis using data from the Tennessee STAR (Student/Teacher Achievement Ratio) experiment. Project STAR was a four-year longitudinal study of elementary school children in Tennessee between 1985 and 1989. At the beginning of the experiment, 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. 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. Actual class size deviated somewhat from these targets due to schoolroom capacity constraints and attrition from participating schools. Figure 2 shows that, in kindergarten, the small ("treated") classes ranged from twelve to seventeen students while the regular ("control") classes ranged from sixteen to twenty-seven students. In addition, teachers were randomly assigned to the classes they would teach. Details of the randomization procedure are provided in Krueger (1999).

The experimental design called for students to remain in the same class type through the third grade. However, for a variety of reasons including sample attrition, behavioral issues, and parental complaints, many students initially randomized into a specific class type eventually dropped out of the experiment altogether or attended another class type for at least one year of the four-year project. Of the 1,900 children initially randomized into a small class in kindergarten, only 857 (45%) attended a small class for all four years. For this reason, we analyze the effects of class size on academic performance in kindergarten only, before any potentially endogenous reallocations to different class types may have taken

Figure 2: The Distribution of Class Sizes



The figure plots the number of kindergarten classes of each size for regular and small classes in the Tennessee STAR experiment

place. In addition, past 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).

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 SATs use item response theory to facilitate comparisons across students and across years for the same student. Our dependent variable is an average of scores on these three exams. For the 2.3% of students with scores on at least one, but not all three, exams, we construct their mean score as the average of the available scores.

We depart from the analysis in Krueger (1999) in using raw scores rather than percentile ranks as the outcome variable. First, 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.

Project STAR collected additional demographic information on the students and their teachers. These include information on student race, gender, free lunch status, absences from school, as well as teacher demographic information and qualifications. Prior research has shown that the treatment and control groups in kindergarten do not differ by these observable characteristics, suggesting that the initial randomization was not compromised (Krueger 1999).

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 (n=3813).

Summary statistics for the estimation sample are in Table 1. The exams, which are 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.

6 Results

In this section we present the estimates of the marginal effects of class size on test scores. The outcome we consider is a simple average across three exams administered to the children at the end of the first year of the experiment. In order to estimate these marginal effects, we first group schools together based on the experimental effects and the class size formation model as documented in section 3.

Table 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.

The STAR experiment randomized students into either treatment or control classrooms at the school level, so before running the clustering algorithm, we first subtract from the dependent and independent variables their school-specific means, thereby partialing out the school fixed effects. We also include binary controls for child gender, race (an indicator for being neither white nor Asian), and free lunch status. The EAMP algorithm takes as given the number of groups. When choosing the group size, the modeler wants to balance the competing attractions of parsimony and allowing for richer patterns of heterogeneity. The EAMP framework allows for as many models relating generating test scores as there are groups. 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 Akaike Information Criterion to select the number of groups. AIC selects four groups as optimal.

We present estimation results for the test score model with four groups in Table 2. The table shows the marginal effects of class size on test scores for each group.

Table 2: Class Size Marginal Effects by group

Group	Class Size Effect	S.E.	Schools	Students
1	-0.042	0.033	18	964
2	0.113	0.026	18	917
3	-0.322	0.023	23	1131
4	-0.038	0.021	20	801
Avg. Effect	-0.087	0.013		
N	3813			

The table shows the results from a regression of test scores on class size interacted with estimated group membership, where schools are clustered into groups according to the model described in section 3. Regressions include 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. The model's estimates for the equation (7) are shown in Table A-1.

We obtain these effects by running a regression of y_{isk} on n_{isk} (along with various controls) within each group. We find striking differences in the marginal effects of class size on test scores across groups. Group two, with eighteen schools, is characterized by a positive marginal effect of class size on test scores, while the other three groups feature a negative marginal effect. The only statistically significant negative effect of class size on test scores occurs at schools in group three. A randomly selected student from one of these schools would be expected to perform 0.322 points better on the exam if her class size were reduced by one. This is nearly a 0.1 sd improvement (the sd of the outcome variable is 3.59). The academic performance of students in group three is therefore highly sensitive to changes in class size. Averaged across all groups and all students, the marginal effect of a one-student reduction in class size is an improvement of 0.087 points, or 0.024 sd, in the outcome measure.

Table 3: Treatment Class Size Support by Group

Group	Class Size (Number of Students)					
	12	13	14	15	16	17
1	0	0.22	0.17	0.14	0.33	0.14
2	0.14	0.10	0.13	0.22	0.23	0.18
3	0.03	0.10	0.25	0.10	0.28	0.24
4	0.07	0.22	0.11	0.28	0.14	0.18

The table shows the Dirichlet prior means for each treatment class size by group. Each reported number can be interpreted as the fraction of observations in each cell. Zeros indicate that a particular group does not generate classes of that size.

In addition to finding heterogenous effects of class size on test scores, we also find that schools responded to the randomization protocol in heterogeneous ways when determining the sizes of treatment and control classes. Tables 3 and 4 display the estimated Dirichlet prior means by group for the class size generation model. In our context, the Dirichlet prior mean corresponds to a prior on the fraction of classes with each size. A higher value, all else equal, indicates that schools within that group are more likely to create a classroom of that size.

Group one chooses relatively large treatment and control class sizes. Group two, the only group with a positive marginal effect of class sizes on test scores, instead, tends to choose uniformly distributed treatment class sizes and large control classes. Group three is similar, while group four is unique in that it only chooses control class sizes of 17-19 and 23-25.

The groupings selected by the EAMP algorithm represent marked patterns of heterogeneity within the data. In Table 5 we show how students in different groups differ by observable characteristics. We repeat the same analysis by group for teacher-specific variables in Table 6 and for school-specific variables in Table 7. The tables show that group three, which has the largest marginal effects of class size on student performance, has the highest fraction of nonwhite students

Table 4: Control Class Size Support by Group

Group	Class Size (Number of Students)											
	16	17	18	19	20	21	22	23	24	25	26	27
1	0	0.07	0	0.11	0.07	0.11	0.13	0.25	0.11	0.07	0.03	0.03
2	0.04	0	0	0	0.09	0.04	0.41	0.09	0.23	0.09	0	0
3	0	0.04	0	0	0.04	0.35	0.25	0.07	0.11	0.04	0.11	0
4	0	0.05	0.05	0.20	0	0	0	0.25	0.40	0.05	0	0

The table shows the Dirichlet prior means for each control class size by group. Each reported number can be interpreted as the fraction of observations in each cell. Zeros indicate that a particular group does not generate classes of that size.

and the most disadvantaged students as measured by free lunch status. Students in group three also attend school for the fewest days of the year, a result due to both absences and the length of the school year. Group two, which is the only group with a positive relationship between class size and student performance, is by contrast the least nonwhite and consists of fewer students qualifying for free lunch. Students attending schools in group two spend an additional two weeks in school relative to students in group three. Table 6 shows that teachers at schools in group three are less likely to have a masters degree than teachers at schools in groups two and four, and also have somewhat less experience than teachers in groups one and two. Group two, in which class size and test scores are positively related, has very few nonwhite teachers.

Table 7 shows that schools in group three are mostly located in inner city or rural settings. The inner city schools in group three are perfectly segregated – 100% of each cohort of kindergarten students in these schools are nonwhite. This suggests that one fruitful way of targeting resources to deliver large returns is to reduce class sizes at highly segregated inner city schools. Altogether, the results suggest that the population of schoolchildren in Tennessee who benefit the most from class size reductions are disadvantaged according to a number of metrics, including race, family income, and the extent of segregation. The findings presented here buttress the subgroup analysis in Krueger (1999) that found that

Table 5: Student Characteristics by Group

Group	Female	Nonwhite	Free lunch	Avg. Days Present
1	0.50	0.39	0.43	154.6
2	0.49	0.17	0.47	162.2
3	0.49	0.46	0.58	152.7
4	0.46	0.26	0.41	157.2

The table shows the average student characteristics by group. Avg. Days Present is determined by both the length of the school year and student absences.

Black students and students qualifying for a free lunch responded more to the reduction in class sizes than did their white and wealthier peers.

In sum, the EAMP algorithm identifies patterns of heterogeneity in who benefits from class size reductions and what types of schools reduced class sizes the most in the Tennessee STAR experiment. The students who benefited the most from the intervention were more likely to be nonwhite students attending segregated schools and to be at an initial disadvantage in terms of family resources.

6.1 Comparison with Other Methods

In Table 8, we compare our estimated model to the linear model estimated via instrumental variable estimator. Both models produce very similar estimates for the nonrandom coefficients on the Female, Nonwhite, and Free Lunch binary variables. The EAMP estimate for class size is smaller in magnitude than the IV estimate. In line with our previous discussion of the IV weighting scheme, we interpret this as evidence of bias due to heterogeneous compliance responses of schools to the experimental variation. In particular, Table 9 shows the average class size reduction by grouped schools. Although the differences do not seem too large, they are enough to generate differential conclusions regarding the average impact of class size reduction in the population.

Table 6: Teacher Characteristics by Group

Group	Masters Degree	Experience	Nonwhite
1	0.25	9.94	0.23
2	0.39	9.48	0.06
3	0.34	9.09	0.17
4	0.43	8.57	0.19

The table shows average teacher characteristics by group. Experience is measured in years.

Table 7: School Characteristics by Group

Group	Inner City	Rural	Suburban	Urban	Avg. Cohort Size
1	3	7	6	2	82.33
2	1	12	2	3	85.56
3	7	11	4	1	84.7
4	5	8	6	1	67.75

The table shows counts of the type of neighborhood where each school is located as well as the average cohort size by group.

Table 8: Model Comparison

	IV	EAMP
Class Size	-0.101	-0.087
	(0.014)	(0.013)
Female	0.650	0.608
	(0.100)	(0.097)
Nonwhite	-1.057	-1.095
	(0.197)	(0.191)
Free Lunch	-1.776	-1.734
	(0.118)	(0.114)
First Stage F-Statistic	15384.26	

The table presents a side-by-side comparison of the three models applied to the data from the Tennesse STAR experiment. The first column presents the Ordinary Least Squares estimate for the effect of class size on test scores. The second column presents the estimates from a two-stage least squares regression that uses the binary randomization into treatment as the excluded instrument. Estimates from our model are in the third column. Heteroskedasticity-robust standard errors are in parentheses.

7 Alternative Specifications

In this section we examine whether the relationship between class size and test scores may be nonlinear. To do this, we estimate nonparametric regressions of test scores on class size separately by group. We first partial out the group-specific means from both test scores and class sizes for the six cells created by $Female \times Nonwhite \times Freelunch$ before estimating the nonparametric relationship between the demeaned test score and the demeaned class size using local linear regression. We use an Epanechnikov kernel and a bandwidth of 7.5 students,

Table 9: Intensity of Class Size Reductions

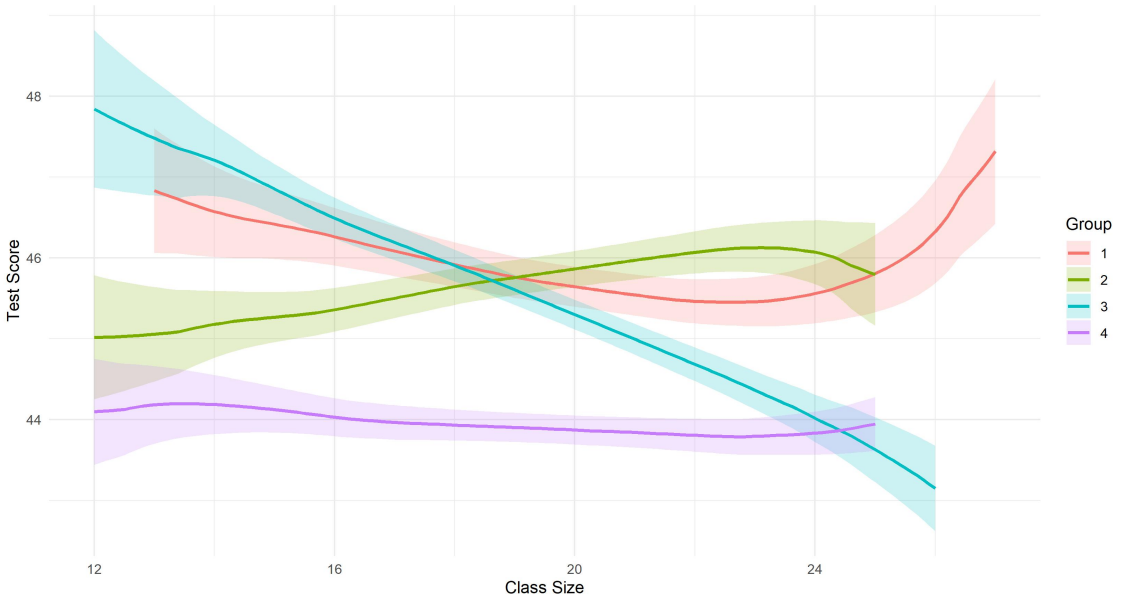
Group	Class Size Effect	Class Size Reduction
1	-0.042	-7.16
2	0.113	-7.37
3	-0.322	-7.01
4	-0.038	-7.47

The table shows the intensity of class size reductions at each group of schools alongside the class size effects from Table 2. Class size reduction is computed as the difference in size between the average treatment class and the average control class in each group.

which lets in about 50% of the data. As the EAMP algorithm clusters schools into groups on the basis of both the distribution of class sizes within each school and the school-specific difference between treatment and control class test scores, the clustering algorithm is agnostic as to the specific relationship relating class sizes to test scores. This approach is entirely consistent with a potentially nonlinear relationship between class sizes and test scores, and the results presented in section 6 should be interpreted as estimates of the best linear approximation to the relationship between class size and test scores by group.

Figure 3 displays the results of the nonparametric regressions of test scores on class size. The figure plots the regressions only over the support of class sizes within each group. The shaded regions represent asymptotic 95% confidence intervals. The figure reveals that group three has a sharp downward relationship between class size and test scores that is approximately linear. The difference in class size between group three students in the smallest classes, of twelve students, and those in the largest classes, of twenty-six students, is approximately 1.25 sd on the exam. This relationship dwarfs the effects seen in other groups. Group four has a nearly flat effect of class size on test scores, while group two has a concave upward-sloping relationship, and group one has a convex rela-

Figure 3: Treatment Intensity and Random Effects



tionship. These nonparametric regressions demonstrate the heterogeneity in the model for test scores across groups and suggest that, when evaluating the effects of interventions that reduce class size at schools that belong to group three, a linear model is a good approximation.

8 Counterfactuals

In this section, we use our model to extrapolate from the actual Tennessee STAR experiment and answer questions of how different, unimplemented versions of the experiment might have influenced student test scores. We use the linear model with $K = 4$ groups to conduct all counterfactual analyses.

Throughout, we have assumed that the coefficients determining the treatment effects, β_{isk}, represent a student-to-school match. Therefore, we analyze counterfactuals that do not reallocate students across schools, but rather preserve the same student-to-school match. These counterfactuals treat the coefficients on class size from Table 2 as fixed, but alter parameters of the Dirichlet distribution that determine class size in each school. To obtain counterfactual policy estimates, we simulate the model 500 times for a particular counterfactual and

average the resulting treatment effects.

The first counterfactual imposes a uniform distribution on the Dirichlet parameters, so that $(\eta_g^{(t)}, \eta_g^{(c)}) = (\eta_h^{(t)}, \eta_h^{(c)})$ for all groups g and h. $\eta_g^{(t)}$ will itself be uniform over the observed distribution of treatment class sizes and $\eta_g^{(c)}$ will likewise be uniform over the observed distribution of control group class sizes. After simulating class sizes from this distribution, we compute the ATE as if the experiment had been conducted with a uniform class size distribution.

The second counterfactual recovers the maximum ATE that could have been achieved by the Tennessee STAR experiment. To do this, we modify the Dirichlet priors to maximize the difference between treatment and control class sizes in all schools with negative marginal returns to class size (those in groups one, three, and four) and minimize this difference for schools (in group two) with a positive marginal return to class size. This maximum ATE is used to provide context to and evaluate the external validity of the estimate that was found in the actual STAR experiment.

The first two counterfactuals make no assumption regarding the school fixed effects. The ATE is computed using the Law of Iterated Expectations, by first computing the ATE for each school, ATE_s, and then computing a weighted average of all the ATE_s, where the weights are determined by the proportion of the sample in each school.

The third counterfactual analyzes the effect of a marginal reduction in class size across all schools on the Black-white test score gap. The fourth counterfactual repeats this exercise comparing students who qualify for free lunch and those who do not.

The final counterfactual forecasts the effect on test scores of reducing class size by five students for everyone in Tennessee. It then asks, What fraction of this effect could be achieved by a targeted policy that reduces class size by five students for the single group with the largest marginal returns to class size?

Table 10 presents estimates of these counterfactual treatment effects. The first row (Status Quo) presents the average treatment effect achieved by the policy

as it was implemented. The second row instead reveals what the average treatment effect of the Tennessee Star experiment *would have been* had there been no systematic correlation between the marginal returns to class size reductions and the treatment intensity. We see that the experiment would have had a slightly lower average class size, a slightly larger difference between treatment and control classrooms, and a marginally larger average treatment effect (ATE). If, instead, the experiment had been designed in such a way to maximize estimates of the average treatment effect, by creating large differences between treatment and control class sizes in the schools with negative marginal returns of class size on student performance and small class size differences in schools with positive marginal returns, the ATE of the experiment would have near trebled, to 1.75 points on the exam. This scenario would have resulted in an average class size of about 20.6 students per class and a mean class size difference of nearly twelve students between treatment and control classrooms. The findings of this counterfactual treatment effect underscore the potentially large gains to reducing class size for students that benefit from these reductions. Recall, from Table 1, that the standard deviation on the outcome variable is 3.59 points. Hence, an experiment that caused an average difference of nearly twelve students between treatment and control classes would have caused a near one-half standard deviation difference in test scores between the two groups.

The remaining three counterfactuals in Table 10 consider a reduction in class size of five students per classroom at all schools in the Tennessee STAR experiment. In the first, we show the Black-white test score gap and the reduction in this gap induced by the counterfactual in both raw scores and percentage terms. While this counterfactual benefits all students in the state due to its nontargeted nature, it still has a nontrivial effect on the test score gap: A five-student reduction in class sizes induces a 24% reduction in this gap. The effects of this policy on the gap in test scores between students qualifying for a free lunch and those who do not is less impressive but not negligible: A five-unit reduction in class sizes induces an 8.2% reduction.

The final panel of Table 10 documents the percentage change in test scores that could have been achieved had the five-unit reduction in class size been targeted

to only one group. The first column shows that the effect of reducing class size for everyone increases average test scores by 0.435 points. However, the second column shows that if the reduction had instead been targeted to only group three, average test scores would have increased by even more, 0.478 points. Thus, the entire benefit of the policy could be achieved by targeting the policy only to the 30% of students at schools in group three, which has the largest marginal returns to class size reductions. The second row of the final panel shows that the effect of this targeted policy on the Black-white test score gap would be to reduce it by 0.24 points. This represents 73% of the effect of the untargeted policy on the Black-white test score gap in the second panel ($\frac{-0.24}{-033} = 0.73$). The results indicate that targeting class size reductions to the most responsive students would simultaneously generate large test score gains and reduce educational inequality.

Altogether our counterfactual treatment effects highlight the significant distributional consequences of policies that can reduce class size as well as show the importance of accounting for the design of the experiment in interpreting treatment effects. The large variation between the status quo and the experimental design that seeks to maximize the average treatment effect underscores the lack of external validity when using the Tennessee STAR experiment to make predictions of the effects of class size reductions in other settings without accounting for a model of test score determination.

9 Conclusion

This paper builds a novel empirical framework that can be applied to large families of randomized controlled trials that allow for endogenous compliance with respect to treatment intensity. In particular, we reevaluate the effects of class size on student learning generated by the Tennessee STAR experiment, one of the most important educational experiments in the United States. This method is consistent with STAR's experimental design, in which students were randomized into either small or regular-sized classrooms but schools were free to set targets for the desired size of each class type. We estimate a model allowing for grouped selection on treatment and control class sizes using a variation on the EM algorithm, following Adusumilli (2020). After grouping schools together based on

Table 10: Counterfactual Treatment Effects

	ATE	Average Class Size	Mean Class Size Difference
Status Quo	0.60	19.02	-7.28
Uniform Class Sizes	0.62	18.65	-7.35
Max ATE	1.75	20.6	-11.7
	Black-White Gap	Effect on Gap	% Reduction
Class Size Reduced by 5	-1.40	-0.33	23.6%
	Free Lunch Gap	Effect on Gap	% Reduction
Class Size Reduced by 5	-1.98	-0.16	8.2%
	Average Effect	Average Effect if Targeted	% of Total Gain
Class Size Reduced by 5	0.435	0.478	109.8%
		Effect on B-W gap	% of Gain in B-W gap
		-0.24	72.5%

The table shows estimates of Treatment Effects for all the counterfactuals we evaluate. The first row presents statistics for the Tennessee STAR experimented as implemented. The second evaluates the average treatment effect (ATE) of the experiment if the Dirichliet prior for class size had been uniform instead of the values estimated in Tables 3 and 4. The third evaluates the ATE of the experiment if it had been designed to maximize ATE. The second and third panels present the effect of reducing class size by five students for everyone on the black-white and free-lunch test score gaps. The final panel shows the how much of the benefits of reducing class size could be achieved by targeting the group with the largest marginal effects of class size on performance. All counterfactuals are estimated without making an assumption on school-specific exam means (fixed effects). Details regarding the construction of these counterfactuals are presented in section 8.

the reduced form of the experiment and the targeted class sizes, we estimate both linear and nonparametric regressions of test scores on class size for each group.

Both our linear and nonparametric results show that nearly all of the treatment effects generated by the Tennessee STAR experiment occurred in one group of schools, constituting just 30% of the schools in the experiment. Students in these schools, which disproportionately enroll low-income and nonwhite students, are very sensitive to changes in class size, with a one-student reduction in class size generating a nearly 0.1 sd increase in test scores. For the remaining schools in the experiment, changes in class size have on average no effect or even a small positive effect on test scores.

We see our research as contributing to the debate on universal versus targeted programs. Our counterfactual analysis shows that a well-targeted investment to reduce class size can generate the same overall effects on test scores as a universal program. The same program would also reduce the black-white test score gap by 72.5% as much as a universal program. Targeted interventions have other benefits not studied in this paper, most importantly that they can be scaled more easily while retaining quality (List 2022).

Moreover, we shed light on the contrasting conclusions from previous studies on class size effects on children's learning. For example, while Hanushek 1997 and Hoxby 2000 argue that class size has null or small impacts on children's learning, other studies that focused on the evaluation of STAR experiment argue for large effects of class size reduction (see for example Krueger 1999; Schanzenbach 2006). Our results suggest that the heterogeneity in class size effects may reflect how different empirical approaches either isolate or average over multiple distinct subpopulations with sharply different responses to class size.

We believe that more research is needed to understand why learning in certain contexts responds significantly to changes in class size, while class size has negligible effects in other contexts. However, our findings suggest that, if policymakers want to invest resources to generate gains in achievement, targeting funds (instead of universal interventions) towards reducing class size in segregated inner city schools is a good place to start.

References

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

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 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.

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.

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa. 2017. "Discretizing unobserved heterogeneity." *University of Chicago, Becker Friedman Institute for Economics Working Paper*, no. 2019-16.

Bonhomme, Stéphane, and Elena Manresa. 2015. "Grouped patterns of heterogeneity in panel data." *Econometrica* 83 (3): 1147–1184.

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.

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.

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.

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

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, 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.

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.

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.

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.

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

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

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*.

A Details of the EAMP algorithm

In this section we describe the various steps of the EAMP algorithm for computing the GRE estimates.

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 $p(\boldsymbol{y}_s, \boldsymbol{n}_s^{(c)}, \boldsymbol{n}_s^{(t)} | \boldsymbol{\alpha}_s, \boldsymbol{x}, \boldsymbol{T}, \theta)$ 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

$$\begin{aligned} \Sigma_{isk} &\leftarrow \left(\Sigma_k^{-1} + Treat_{isk} Treat_{isk}^{\intercal}\right)^{-1}, \\ \mu_{isk} &\leftarrow \Sigma_{isk} \left(\Sigma_k^{-1} \mu_k + Treat_{isk}(y_{isk} - x_{isk}' \theta)\right) \end{aligned} \tag{A-1}$$

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 into one of the K groups by maximizing (14) with respect to $w_s(k)$. As in Adusumilli (2020), group assignments are obtained as the solution to the following problem:

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

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

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

which can be grouped into three separate terms:

$$\begin{aligned}
\ln \frac{p(\boldsymbol{y}_s | \boldsymbol{\beta}_s, \boldsymbol{x}, \boldsymbol{T}, \theta) \pi(\boldsymbol{\beta}_s | \boldsymbol{\rho_k})}{q_{sk}(\boldsymbol{\beta}_s)} &= -\frac{1}{2} \sum_{i:s(i)=s} \left\{ (y_{isk} - x_{isk}' \theta)^2 + \mu_k^\top \Sigma_k^{-1} \mu_k - \mu_{ksi}^\top \Sigma_{ksi}^{-1} \mu_{ksi} \right\} \\
&\quad + \frac{1}{2} \sum_{i:s(i)=s} \left\{ \ln |\Sigma_{ksi}| - \ln |\Sigma_k| \right\} + \text{ const} \qquad \text{(A-4)} \\
\ln \frac{p(\boldsymbol{n}_s^{(c)} | \boldsymbol{p}_s^{(c)}) \pi(\boldsymbol{p}_s^{(c)} | \boldsymbol{\eta}_k^{(c)})}{q_{sk}(\boldsymbol{p}_s^{(c)})} &= \ln \frac{N_s^{(c)}!}{\prod_{j=1}^J n_{skj}^{(c)}!} \frac{B(\eta_k^{(c)} + n_{sk}^{(c)})}{B(\eta_k^{(c)})} \\
&= \ln B(\eta_k^{(c)} + n_{sk}^{(c)}) - \ln B(\eta_k^{(c)}) + \text{const} \qquad \text{(A-5)} \\
\ln \frac{p(\boldsymbol{n}_s^{(t)} | \boldsymbol{p}_s^{(t)}) \pi(\boldsymbol{p}_s^{(t)} | \boldsymbol{\eta}_k^{(t)})}{q_{sk}(\boldsymbol{p}_s^{(t)})} &= \ln \frac{N_s^{(t)}!}{\prod_{j=1}^J n_{skj}^{(t)}!} \frac{B(\eta_k^{(t)} + n_{sk}^{(t)})}{B(\eta_k^{(t)})} \\
&= \ln B(\eta_k^{(t)} + n_{sk}^{(t)}) - \ln B(\eta_k^{(t)}) + \text{const} \qquad \text{(A-6)}
\end{aligned}$$

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 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, by solving

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

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).

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:

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

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

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 M equations in M unknowns for control group k is given by

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

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:

$$\psi\left(\tilde{\eta}_{kl}^{(t)}\right) - \psi\left(\sum_{l=1}^{L}\tilde{\eta}_{kl}^{(t)}\right) = \frac{1}{N_{sk}}\sum_{s} w_s(k)\left\{\psi\left(\eta_{skl}^{(t)}\right) - \psi\left(\sum_{l=1}^{L}\eta_{skl}^{(t)}\right)\right\}, \text{ for each } l = 1, \ldots, L \tag{A-12}$$

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, $k = 1, \ldots, K$.

Table A-1: Reduced Form Treatment Effects by Group

Group	Reduced Form Effect	S.D.	Schools	Students
1	0.540	8.276	18	964
2	-0.706	5.883	18	917
3	2.409	5.571	23	1131
4	0.201	4.111	20	801
Avg. Effect	0.723	6.308		
N	3813			

The table shows the estimates from the EAMP algorithm of the test score equation (7). The algorithm clusters schools into groups on the basis of the reduced form experimental effects and the class size model. A positive reduced form effect indicates that the randomization caused smaller classes to achieve higher scores than regular-sized classes in the experiment.