Heterogenous Treatment Effects in Early Language Literacy Interventions

A Preprint

Alexander Chin*
alexanderchin@college.harvard.edu

Asher Spector*
asherspector@college.harvard.edu

December 2019

Abstract

Early language literacy (ELL) interventions are known to improve educational outcomes on average [3]. However, policymakers facing budgetary trade-offs may wish to target interventions towards teachers and students who will benefit the most. Unfortunately, current studies analyzing heterogeneous treatment effects for ELL interventions lack rigor: they rely on heuristic cross-study comparisons or ad-hoc subgroup analysis.

We apply sparse LASSO techniques to rigorously identify heterogeneous effects of a randomized intervention in Miami [7]. We find that classrooms with lower support for language-learning gain more from these interventions. Moreover, we design a budget-constrained individualized treatment rule (ITR) for policymakers, and demonstrate that our ITR is more efficient than a competitor constructed from the relevant literature [6]. Our main results are cluster-robust and do not rely on asymptotic theory. This contributes to the literature by rigorously demonstrating that ELL interventions equalize instructional quality among classrooms (139 words).

1 Introduction

1.1 Motivation

According to the National Assessment of Educational Progress, 37 percent of fourth graders cannot complete basic reading requirements [14]. This has motivated a substantial amount of government spending on programs aimed at

remedying poor reading ability starting in preschools, such as the Head Start preschool program that cost 6 billion dollars annually [14]. Many of these programs acknowledge the fact that most childcare professionals lack formal training and secondary education, and aim to improve student outcomes by training teachers to implement curriculums which facilitate their student's language and literacy development [5, 9, 12, 15]. We refer to these programs as *early language literacy interventions* (ELL interventions).

Unfortunately, policymakers always face budgetary trade-offs, and may not have the resources to intervene in every single school. Thankfully, researchers can help mitigate this issue by identifying groups of students and teachers who stand to gain the most from ELL interventions, allowing policymakers to target interventions towards these groups and maximize the positive impact of ELL programs.

Our paper aims to help policymakers identify these groups in two ways. Using data from a randomized-control trial of ELL interventions in Miami-Dade, we will first analyze whether ELL interventions have heterogenous effects on instructional quality and interpret the policy implications of these heterogenous effects. Second, we will use our findings to design an individualized treatment rule (ITR) which helps policymakers appropriately target ELL interventions when facing budget constraints.

1.2 Findings and Contributions to the Literature

This paper has two main findings. First and most importantly, we show that ELL interventions have stronger positive effects on instructors who score poorly on baseline measures of instructional quality. We interpret this as strong evidence that ELL interventions equalize instructional quality between classrooms. Second, we construct an individualized treatment rule which, given a budget constraint, automatically prioritizes certain classrooms to receive treatment. We demonstrate rigorously that for a given budget constraint, this treatment rule improves instructional quality more than both a randomized treatment rule and a competitor treatment rule constructed based on a literature review. From this we conclude that our estimates of heterogenous treatment effects (and the methodology behind it) can substantially improve the efficiency of budget-constrainted ELL interventions.

Our paper offers both a substantive and a methodological contribution to the literature. Substantively, very little existing literature has analyzed heterogeneous effects of ELL interventions on *teacher behavior* and instructional quality. In contrast, most of the heterogenous research on ELL interventions analyzes the effect of reading interventions on student test scores through meta-analysis [10, 1]. However, evaluating how interventions affect instructional quality is crucial to understanding the success of interventions, since a teacher-training program which fails to change teacher behavior

will almost certainly fail to improve student outcomes. Indeed, previous research in other contexts (often in high schools) has found counterintuitive and even negative effects of teacher-training interventions, especially on veteran teachers, who resent and resist top-down policy changes [17]. As a result, we think this question genuinely matters to policymakers.

Note that one other paper that we know of does explicitly evaluate heterogenous effects of ELL interventions on teachers. In particular, Layzer et al. in 2007 [9] analyze the same dataset that we use in this paper. Although they largely focus on computing average treatment effects, they find that ELL interventions have slightly larger positive impacts for Spanish-speaking teachers. Our result generalizes theirs, as we find that regardless of the underlying cause for low performance, ELL interventions have an equalizing effect for these underperforming teachers. Our finding is substantially more actionable for policymakers who want to improve teacher quality, as it implies they ought to target all underperforming teachers, not just Spanish-speaking ones.

This paper's second contribution is methodological. Most literature analyzing the heterogeneous outcomes of literacy interventions is through meta-studies, and these rely on cross-study comparison of effect sizes [10, 1]. Unfortunately, meta-analysis assumes a high level of comparability between studies of vastly different interventions. As a result, perhaps unsurprisingly, such research has found contradictory results on whether ELL interventions equalize outcomes between students or exacerbate inequality by helping higher-income students more [1, 19, 16].

All other studies of heterogenous treatment effects in this field, to our knowledge, rely on ad-hoc subset analyses, where researchers split the data into two arbitrary groups and compare average treatment effects between the groups. While this analysis is not always incorrect, researchers usually are unsure exactly *which* pre-treatment covariates influence treatment effects, and therefore run many different subset analyses, leading to multiple testing problems. For example, Layzer et al. present 22 different regressions and do not seem to correct for multiplicity [9].

In contrast, this paper specifically uses post-selective LASSO methods pioneered in [8] to select relevant covariates and minimize the effect of multiplicity. Even though we do not a priori know which covariates interact with the treatment, we are able to automatically detect these covariates and generate exact p-values and confidence intervals for a small set of selected covariates (which are then robust to multiple testing). Additionally, our LASSO method can estimate the treatment effect for each individual. This allows us to construct individualized treatment rules which are substantially more efficient than individualized treatment rules based on coarser subgroup analyses, which only estimate the mean treatment effect for different subgroups.

2 Research Questions and Hypotheses

This paper answers three main research questions.

1. Do ELL interventions have highly heterogenous causal effects on teacher instructional quality?
2. If so, do ELL interventions equalize instructional quality across classrooms?
3. Can we design individualized treatment rules to make ELL interventions more efficient?

These research questions are slightly vague by design, since a priori we make no formal assumptions about *which* covariates make a treatment more or less likely to be effective. Instead, we hope to automatically identify factors which affect treatment efficacy in a principled, statistically rigorous fashion, as we will discuss in Section 4.

That said, we hypothesize initially that the answer to all of these questions is *yes*. In particular, we suspect that ELL interventions do have differential effects on different teachers' instruction quality, and that ELL interventions are more effective for weaker teachers. Therefore, we expect that targeting ELL interventions to underperforming teachers should improve efficiency.

3 Dataset and Context

In this section, we first offer a high level overview of the Project Upgrade experiment and justify our focus on instructional quality as the outcome of interest.

3.1 Context and Experimental Design

In this paper, we analyze data from a two-year randomized control trial (RCT) called Project Upgrade which took place in Miami-Dade, Florida between 2003 and 2009 [12]. Florida's Early Learning Coalition commissioned Project Upgrade due to concerns that some pre-school teachers lacked training, which hindered preschool-age children's language development, especially children receiving childcare subsidies. As a result, Project Upgrade tested a teacher-level intervention intended to improve the quality of literacy-related instruction in pre-school classrooms serving lower-income students. In particular, the intervention offered teachers 18 months of professional development opportunities and additionally provided teachers with tools to assess children's progress, as well as the materials and training necessary to implement Florida sanctioned literacy development curriculum. [2]

[2]Technically, Project Upgrade tested three separate treatments, but these three interventions were fairly similar: they included the same professional development opportunities, provided the same literacy materials such as books and assessment tools, and had only

Project Upgrade was completely-randomized at the (center) level, with $n = 162$ centers participating and a control group of size $n_0 = 54$. Note that the experiment provided the same physical classroom materials (e.g. books) to the control group, but only the treatment group received instructional training. The designers also employed 18 randomization blocks based on pre-treatment covariates to ensure covariate balance between the treatment and control groups: these pre-treatment covariates included teacher education levels, student demographics, and more.

3.2 Response Variables

From 2003 to 2005, the study measured 2 types of outcomes. First, in 2005, the study recorded the scores of students on standardized tests designed to measure language literacy. Second, in late 2004 and early 2005, researchers visited teachers' classrooms and rated the quality of instruction pertaining to literacy development using a methodology known as OMLIT [2]. The OMLIT rates teachers along four relevant axes: "support for print knowledge" (e.g. alphabet knowledge, sound-letter correspondence), "support for print motivation" (motivating children to read), "phonological awareness" (e.g. breaking words apart into syllables), and "oral language" (giving children practice speaking in English).

In this paper, we focus on the *second* set of observations, the OMLIT observations from 2004, to estimate heterogenous treatment effects on instructional quality. We focus on the OMLIT responses because students often switched between classrooms within the same center (and therefore switched between teachers) between 2003 and 2005. Although this does not pose a problem for estimating average treatment effects, since every classroom in each center is assigned to the same treatment group, this makes it difficult to tease apart whether the treatment has differential impacts on different teachers using student-level data. In contrast, the OMLIT observations (i) take place at the end of the school year in 2004, before any classroom switching takes places, and (ii) measure the direct impacts on teacher behavior, which is more directly related to our hypothesis.

Although policymakers and researchers might reasonably wonder whether the OMLIT observations measure anything useful, repeated analyses have shown that OMLIT scores are strongly related with student outcomes, and OMLIT observations are well-established in the language literacy literature as a measure of instructional quality (see [2], [20], [9]). Additionally, from a substantive perspective, OMLIT scores reward types of instruction which are fairly uncontroversially beneficial, such as encouraging children to read and working with struggling students individually,

slightly different curricula. As a result, the initial survey authors grouped the three treatments and considered them as one "treated group" to increase effective sample size: we will do this as well (see [9]).

etc [12]. Finally, since many ELL interventions aim to improve student outcomes via improving instructional quality, improving teacher performance on measures like the OMLIT is likely a pre-requisite to improving student outcomes.

Throughout the paper, we standardize all four OMLIT observations so the empirical mean and variance in the datasets are respectively 0 and 1.

4 Statistical Methodology

4.1 Estimands of Interest

For each observation i, let $\mathbf{X}_i \in \mathcal{X}$ be a vector of pre-treatment covariates, where $\mathcal{X}$ is the support of the covariates, and let $Y_i(0), Y_i(1) \in \mathbb{R}$ be the potential outcomes of one of the OMLIT observation types. For each OMLIT observation, we have two types of estimands. First, we seek the conditional average treatment effect (CATE), defined as

$$\tau(x) = \mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x]$$

Second, we would also like to quantitatively analyze whether our model of heterogenous effects can benefit policymakers. To do this, given an estimator $\hat{\tau}(x)$ and "budget" p which represents the maximum proportion of individuals we may give the treatment to, we will design an individualized treatment rule (ITR) $f : \mathcal{X} \to \{0, 1\}$ where $f(\mathbf{X}_i) = 1$ indicates assigning unit i the treatment, and $P(f(\mathbf{X}_i) = 1) = p$. To evaluate the efficacy of this treatment rule, we will estimate the Population Average Prescriptive Effect (PAPE), introduced by [6], which compares the average outcome under an ITR to a treatment rule which randomly assigns individuals the treatment with probability p. Formally, we denote

$$\tau_f = \mathbb{E}[Y_i(f(\mathbf{X_i}) - pY_i(1) - (1 - p)Y_i(0)]$$

Additionally, we would like to compare our methods to those of the literature. Although to our knowledge no other study has created an ITR based on ELL intervention data, we apply the results of [9], who analyze the dataset, to formulate a budget-constrained competitor individual treatment rule g. Using this g, we calculate the Population Average Prescriptive Effect Difference (PAPD), which intuitively measures the "efficacy gap" between f and g:

$$\Delta_p(f, g) = \tau_f - \tau_g$$

4.2 Causal Identification Assumptions

We now lay out and justify the assumptions we need to *identify* our parameters of interest.

Assumption 4.1 (No Interference)**.** *The outcome of each $Y_i(T)$ does not depend on the treatment of the other centers.*

This assumption is likely to hold for two reasons. First, randomization was clustered at the center level, and there are no formal interactions between the centers [9]. Second, our treatment effect is measured over the course of a single school year, so no individuals and teachers switched schools during this time (our dataset confirms this).

Assumption 4.2 (Random Sampling)**.** *For each of the n centers, we assume that* $Y_i(0), Y_i(1), \mathbf{X}_i$ *are independently sampled from some overall population.*

In this case, the “overall population" is the set of centers willing to participate in the study.

Assumption 4.3 (Nondegeneracy)**.** *For each i,* $0 < \mathbb{P}(T_i = 1|Y_i(1), Y_i(0), \mathbf{X}_i) < 1$

This experiment was completely randomized and satisfies this assumption by design.

Assumption 4.4 (Unconfoundedness)**.** *For each i,* $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}$

Since this is a randomized control experiment, this assumption should hold by design. However, we check the pre-treatment covariate balance just in case, and find that they are balanced well (see Figure 1–note the percentage of Spanish speaking teachers is normalized). Note that no centers dropped out of the study upon receiving their treatment, and only 2 centers dropped out over the course of the first year, so attrition poses no threat to this assumption either.

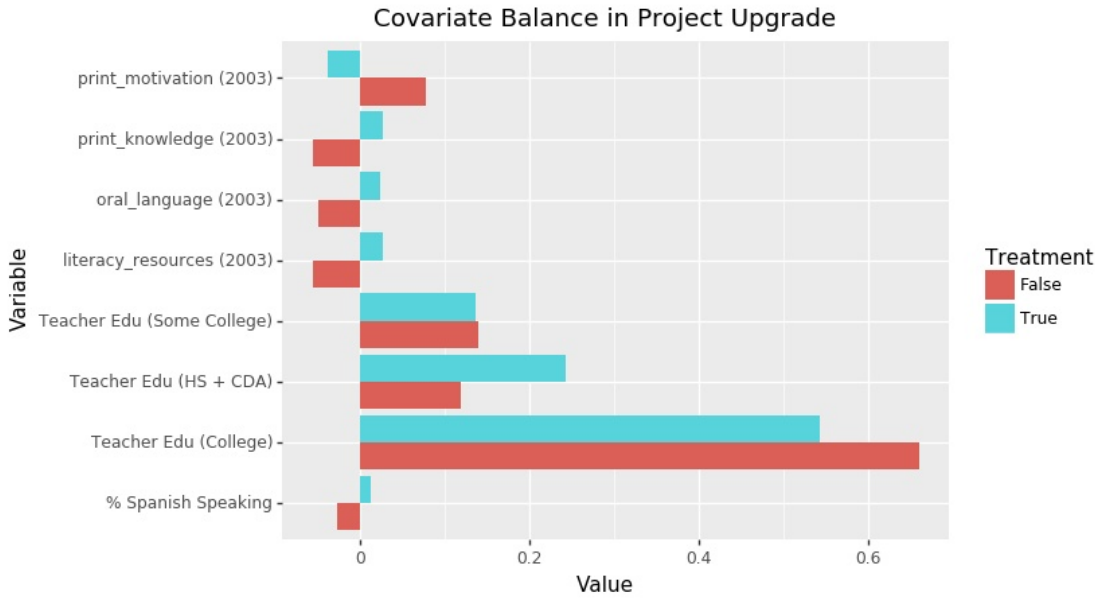


Figure 1: Pre-Treatment Covariate Balance in Project Upgrade, Select Covariates

Under these assumptions, both the PAPE and the CATE are estimable (see [7, 6]).

4.3 Estimation and Inference Strategy, Part 1: Variable Selection and Post-Selective Inference

Our first goal is to estimate $\tau(x) = \mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x]$. We first construction interaction terms $Z_{ij} = T_i \cdot Z_{ij}$ for each i and j, and then model Y_i as linear function of the covariates plus the interaction terms:

$$Y_i = \mu + \beta^T Z_i + \gamma^T X_i + \epsilon_i \tag{1}$$

where ϵ_i is i.i.d. random noise with mean 0. Note that ϵ_i need not be Gaussian for these parameters to be well defined: we will still be able to conduct valid inference even if the model is mispecified.

This may be difficult to estimate in general, since x may be very high dimensional and we are agnostic a priori as to which variables may have a nonzero interaction with the treatment. For example, in our application, $\mathbf{X}_i$ contains nearly 50 different covariates, including pre-treatment response levels, demographic information, and more. These covariates are also highly correlated, and thus for our sample size ($n = 160$), running a linear regression with 50 covariates plus 50 interaction terms will likely yield strange results.

Instead, we will perform *model selection*, largely following the approach of Imai and Ratkovic 2013 [7]. In particular, we will identify a sparse subset of covariates $\tilde{X}_i \subset X_i$ using the data, and then set the coefficients outside of this selected set are uniformly zero (this follows the approach of [7, 11]). Under this selected model, we will then be able to estimate $\tau(x)$ using a lower-dimensional model:

$$\tau(x) = \tau(\tilde{x}) = \mathbb{E}[Y_i(1)|\tilde{X}_i = \tilde{x}_i] - \mathbb{E}[Y(0)|\tilde{X}_i = \tilde{x}_i]$$

We rely on the LASSO regression pioneered in [21], which estimates $\hat{\beta}$ and $\hat{\gamma}$ by minimizing the penalized likelihood function:

$$\hat{\beta}, \hat{\gamma} = \arg\min_{\beta,\gamma} L(\beta, \gamma; X, Y) = \arg\min_{\beta.\gamma} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\gamma_j|$$

The l_1 penalizations ensure a sparse solution where some coefficients β_j and γ_j are “shrunk" to zero. We pick values for λ_1 and λ_2 via a grid search and cross validation on the mean squared error. In particular, for each possible value of λ_1 and λ_2 in a grid of 64 values, we employ 5-fold cross validation to estimate out-of-sample mean-squared-error $\sum(Y_i - \hat{Y}_i)^2$ for the resulting LASSO model, and pick the λ values with the lowest mean-squared-error. This is similar but not identical to the approach of [7], who use an $L2$-SVM and a different cross-validation statistic. We use the statsmodels package and Sklearn packages [4, 18].

Once we have successfully tuned our hyperparameters λ_1 and λ_2, the lasso estimates $\hat{\beta}$ and $\hat{\gamma}$ define an “active set" of selected variables $\mathcal{B} = \{j : \hat{\beta}_j \neq 0\}$ and $\mathcal{G} = \{j : \hat{\gamma}_j \neq 0\}$. Having selected a sparse set of coefficients, we next regress Y_i on $Z_{i,\mathcal{B}}, X_{i,\mathcal{G}}$ using ordinary linear regression to obtain post-selective coefficients $\hat{\beta}_S$ and $\hat{\gamma}_S$.

We would like to now perform two types of inference. First, we would like to obtain valid confidence intervals for each component of $\hat{\beta}_S$ and $\hat{\gamma}_S$, which can be viewed as "interpretable" components of our CATE $\tau(x)$. Since we have used the data twice, first to select a model and then to fit $\hat{\beta}_S$ and $\hat{\gamma}_S$, we apply post-selective inference adjustments pioneered in [8, 13] to obtain exact confidence intervals and p-values for $\hat{\beta}_S$ and $\hat{\gamma}_S$. These adjustments are fairly complex and beyond the scope of this paper, so we refer the interested reader to the aforementioned papers. We employ the selective inference package in Python for computation.

Second, for any particular $x \in \mathcal{X}$, we would like to obain point and interval estimates $\tau(x)$, the CATE. Under the assumed model, $\tau(x) = \beta_S^T X_{\mathcal{B},i}$, so we may consistently estimate $\hat{\tau}(x) = \hat{\beta}_S^T X_{\mathcal{B},i}$. We employ the bootstrap to obtain standard errors for $\hat{\tau}(x)$ which are asymptotically correct under the assumption that the model is correctly specified.

To summarize, our estimation strategy is as follows. First, we employ 5-fold cross validation across a grid of regularization values for the LASSO, and pick the regularization values which minimize MSE. Second, we fit the LASSO using these optimal regularization values to select variables $\mathcal{B}, \mathcal{G}$. Third, we fit the selected model using ordinary linear regression to obtain coefficients $\hat{\beta}_S, \hat{\gamma}_S$, and apply post-selective techniques developed in [8, 13] to obtain exact confidence intervals for $\hat{\beta}_S$ and $\hat{\gamma}_S$. Finally, for any x, we may calculate point and uncertainty estimates for $\hat{\tau}(x) = \hat{\beta}_S^T x$ by bootstrapping.

4.4 Estimation Strategy, Part 2: Individualized Treatment Rules

Next, we design a budget-constrained individualized treatment rule $f : \mathcal{X} \to \{0, 1\}$. As observed in [6], if a policymaker can treat at most p percent of the population, a natural treatment rule given a CATE estimate is

$$f(X_i) = \mathbb{I}(\hat{\tau}(X_i) > c_p) \text{ where } c_p = \inf\{c \in \mathbb{R} : \mathbb{P}(\hat{\tau}(X_i) > c) \leq p\}$$

where in practice we estimate $\hat{c}_p$ using the empirical quantile of the observed data. A policymaker might naturally be interested in quantifying how effective this ITR is: to answer this question, we compare f to two alternatives.

First, we compare f to an alternative which randomly assigns treatments to the same proportion p of the population, in particular by estimating the PAPE, defined as $\tau_f = \mathbb{E}[Y_i(f(\mathbf{X_i}) - pY_i(1) - (1-p)Y_i(0)]$. In [6], Imai and Li showed

$$\hat{\tau}_f(\hat{c}_p) = \frac{n}{n-1}\left[\frac{1}{n_1}\sum_{i=1}^{n} Y_i T_i f(X_i) + \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)(1-f(X_i)) - \frac{p}{n_1}\sum_{i=1}^{n} Y_i T_i - \frac{1-p}{n_0}\sum_{i=1}^{n} Y_i(1-T_i)\right]$$

is approximately-unbiased and derived an estimate for its variance.[3] The bias and variance expressions are fairly complicated so we refer the reader to [6].

[3]Note that calculating their probability bound on the bias assumes Lischpitz-continuity of the CATE, which holds under our linear model, and is a fairly common assumption (see [22]).

Second, we would like to compare our model of heterogenous effects to those of the literature. Although to our knowledge no previous analyses of ELL interventions have constructed ITRs, we construct an individualized treatment rule $g : \mathcal{X} \to \{0, 1\}$ based on the limited heterogenous analysis done by Layzer et. al [9], who analyze the same Project Upgrade dataset. Layzer et al. perform (several) subset analyses and ultimately find that classrooms with higher-proportions of Spanish-speaking students benefit more from the intervention. As a result, given a budget p, the treatment rule g selects the proportion p of classrooms with the highest proportion of Spanish-speaking students. We can then estimate the PAPD as defined earlier using the following estimator, which was also proposed by [6]:

$$\hat{\Delta}_p(f, g) = \frac{1}{n} \sum_{i=1}^{n} Y_i T_i (f(\mathbf{X}_i) + g(\mathbf{X_i})) + \frac{1}{n_0} Y_i (1 - T_i)(g(\mathbf{X}_i) - f(\mathbf{X_i}))$$

This estimator is approximately unbiased under the same assumptions as before. Although its variance is unidentifiable, [6] develop a conservative estimate of the variance.

This method will yield biased results if we use the whole dataset to both formulate the treatment rule f and estimate $\hat{\tau}_f(\hat{c}_p)$ and $\hat{\Delta}_p(f, g)$, since the model will overfit the data. Thus, we split the data into two parts. First, we retrain our model for $\hat{\tau}$ using only a randomly selected 60% of the data, and we use the last 40% of the data to estimate the PAPE. It is worth noting that since we are splitting the data and the dataset is so small to begin with, we set our desired significance level 10% *prior* to computing the PAPEs and PAPDs. (As we will see later, it turns out that many of the p-values are below 5% anyway). We evaluate the PAPE with a budget constraint of $p = 1/2$.

We estimate the treatment effect and standard errors by re-implementing the PAPE function in the R experiment package in python. Our code for all estimation and results is publicly available on GitHub. Although we do not have the right to re-publish the dataset, the dataset can be downloaded at [12].

5 Results

5.1 Interaction Term Coefficients

Below we present the estimates and confidence intervals for the selected interaction terms (and some relevant covariates), e.g. the estimates $\hat{\beta}_S$ and $\hat{\gamma}_S$. Recall that we estimate these parameters 4 times, once for each OMLIT/response measurement.

Interestingly, the LASSO selects strikingly similar covariates and interaction terms for all four response variables. In all 4 cases, the LASSO model selects the baseline, pre-treatment version of the response both as a covariate and as an interaction term, as well as the treatment variable itself. Moreover, with one exception (see the appendix), these

	2003_response	2003_response_interaction	Treatment
print_knowledge	0.109	0.059	0.528*
literacy_resources	0.642**	-0.443*	-0.633
oral_language	0.350*	-0.393*	0.596
print_motivation	0.377**	-0.474*	0.599

Table 1: Significant Interaction and Covariate Terms. The table presents effect sizes for the significant variables detected by the post-selective LASSO procedure for each OMLIT response variable measuring instructional quality. The row names indicate the response variable as measured in 2004, one year after treatment (e.g. print knowledge measures one facet of instructional quality). The columns correspond to effect sizes of the 2003 response as a covariate, as an interaction term, and the treatment. ** $p < 0.01$, * $p < 0.05$

characterize all of the significant effects detected by the procedure. As a result, we report these effect sizes below in Table 1. The model does select other features; however, they are not statistically significant.

The coefficients follow the same pattern for three out of the four responses: a higher baseline (2003) level of instructional quality is positively associated with instructional quality in 2004, but the interaction term between baseline instructional quality and post-treatment instructional quality is negative and significant. We interpret this as strong evidence that the intervention is more effective for teachers who initially have lower instructional quality. Moreover, for these three responses, only the interaction term is statistically significant, and the treatment indicator is statistically insignificant. We take this as further evidence that treatment effects are highly heterogenous, and in particular, the equalizing effect of ELL interventions constitutes the majority of the treatment effect.

Additionally, the effect sizes are fairly large: they hover around -0.4 for the interaction terms. To interpret them, note that all variables are standardized. As an example, the effect for the interaction term for print motivation (-0.474) indicates that a one standard deviation increase in a teacher's baseline level of support for print motivation (e.g. encouraging children to read more) *decreases* the causal effect of the treatment by 0.47 standard deviations. Admittedly, it's tricky to interpret what a "one standard deviation" increase in OMLIT scores means. For context, however, for all four OMLIT measures, teachers with college degrees had baseline OMLIT measurements less than 0.3 standard deviations above their high-school educated counterparts. Thus, informally, the interaction term between the treatment and the baseline response level has a causal effect *more negative* than the positive association between a college-education and OMLIT scores.

response	PAPE	lower	upper	SE	pval	PAPD	lower	upper	SE	pval
literacy_resources	0.299**	0.037	0.561	0.134	0.025	-0.292	-0.672	0.088	0.194	0.132
print_knowledge	-0.189	-0.457	0.079	0.137	0.167	0.468**	0.094	0.843	0.191	0.014

Table 2: PAPE and PAPD Estimates. In this table, we present PAPE and PAPD estimates and uncertainty estimates for ITRs for print knowledge and literacy resources. The PAPE estimates compare our ITRs to a randomized treatment rule, whereas the PAPD estimates compare our ITRs to a language-based ITR we derive from the results of [9]. All ITRs have a budget constraint of $p = 0.5$. * p < 0.1, ** p < 0.05, *** p < 0.001

We consider this to be the main positive result of the paper. The full table of selected variables, coefficient sizes, confidence intervals, and more is available in appendix $A.1$ for each response.

5.2 PAPE and PAPD Estimation

Since we retrain the LASSO model for $\hat{\tau}$ on only 60% of the data and our dataset is so small to begin with, two of the four models unfortunately fail to converge to anything useful (i.e. they do not select any interaction terms). In Table 2, we report PAPE and PAPD results for two responses where the LASSO trained properly (literary resources and print knowledge).

Two of the four PAPE/PAPD estimates are statistically significant: the PAPE estimate of 0.3 for the literacy resources response is significant and positive, as is the PAPD estimate for print knowledge of 0.47. We can interpret these quantities as reporting, respectively, that the ITRs derived from the LASSO model improve instructional quality by 0.3 and 0.47 standard deviations with regard to literacy resources and print knowledge when compared to a randomized or language-based treatment rule. Again, these effect sizes are fairly large: in both cases, the effect size of the PAPE/PAPD is larger than the mean difference in instructional quality between college and non-college educated teachers. The PAPD/PAPE estimates overall provide some evidence that our ITR would allow policymakers to more efficiently target ELL interventions.

The other two of the four PAPE/PAPD estimates are negative, but they are statistically insignificant. That said, we admit that holistically that the results from this section look a bit noisy, and as such we regard this set of results as weaker evidence of heterogenous treatment effects than the previous section. However, the noiseiness of these estimates ought to be unsurprising, since unlike the previous section, this analysis required sample-splitting.

In 3b we present a visualization of $\hat{\tau}$. The graph shows the estimated treatment effects for all centers (note that PAPE/PAPD estimates are only calculated with a subset of the data, but it is useful to visualize all the centers). In 3a we see that for the 3 of 4 OMLIT measures with statistically significant interactions terms (literacy resources, oral language, and print motivation), the estimated treatment effect is highly negatively correlated with the baseline OMLIT score from 2003. This indicates that our $\hat{\tau}$ also predicts the equalizing effect of interventions that we found in our previous regression. Although this does not robustly prove anything, it is in agreement with our other findings.

6 Robustness and Model Misspecification

Since the Project Upgrade experiment was randomized, the identification assumptions are highly plausible, as discussed in Section 4.2. As a result, we discuss and evaluate the plausibility of the model defined in equation (1) in Section 4.4. We discuss broader (non-methodological) limitations of our findings in the conclusion.

Recall that we model the OMLIT teacher instructional quality measures as follows:

$$Y_i = \mu + \beta^T Z_i + \gamma^T X_i + \epsilon_i \tag{2}$$

where Z_i are interaction terms, and X_i is a list of about 40 pre-treatment covariates. We perform selective inference based on this model to obtain sparse coefficients $\hat{\beta}_S$ and $\hat{\gamma}_S$ and corresponding uncertainty estimates.

What could go wrong in this estimation procedure? First, it's possible that the errors ϵ_i are not Gaussian. Technically, model inference will still be valid in this case, since the coefficients $\hat{\beta}$ and $\hat{\gamma}$ will converge to the best linear unbiased predictors by the Gauss-Markov Theorem. However, these parameters are certainly more interpretable in the Gaussian case, so we validate in 2a (see appendix) that the model residuals follow an approximately Gaussian distribution.

Second, since the post-selective approach assumes homoscedasticity, our p-values are not heteroskedacicity robust. However, by grouping the data by the quantiles of baseline level of the response and calculating the standard deviations of the residuals, we demonstrate in 2b that the homoscedasticity assumption holds fairly well in our dataset.

7 Conclusions

Overall we find that the ELL intervention has an equalizing effect on teacher effectiveness. The cross-validated LASSO approach selects 2003 baseline OMLIT scores, treatment, and the interaction between 2003 baseline and treatment for all four outcome measures of teacher effectiveness. While the 2003 baseline scores are positively associated with 2004 scores, the interaction between the 2003 baseline OMLIT score and treatment has a statistically significant effect on 3

of the 4 measured OMLIT outcomes (literacy resources, oral language, and print motivation) of about -0.4. These imply the main result of the paper, which is that a teacher with a higher baseline OMLIT score experiences a weaker effect from the intervention. Overall, ELL interventions have an equalizing effect on teacher quality. This is a positive result for policymakers utilizing ELL interventions as a method of closing the gap.

Our analysis is subject to at least two major limitations.

First, our analysis focuses exclusively on the OMLIT measure of instructional quality, as opposed to student test scores. Although this makes sense methodologically (see Section 3.2), the ultimate goal of ELL interventions is to improve student test scores, even if they accomplish this by measuring and improving teacher quality. Thus, it would be worthwhile to extend this analysis to carefully analyze student-level data as a response.

Second, although our methodology is reasonable, we think the individualized treatment rule we constructed can be substantially improved upon. We did find that our ITR is more efficient than competitor rules, and this result was statistically significant. However, we acknowledge that this finding looks noisy and is less persuasive than our primary result, probably because our methodology requires data-splitting, and our dataset is extremely small to begin with ($n = 162$). Thus, we suspect that future research which constructs ITRs based on larger datasets will design substantially better treatment rules.

References

[1] Valentine JC Cooper H Charlton K and Muhlenbruck L. "Making the most of summer school: a meta-analytic and narrative review." In: *Monographs of the Society for Research in Child Development* (2000).

[2] Babette Gutmann Barbara Goodson Adrienne von Glatz Jennifer Hamilton Ann Webber Patricia Troppe David Judkins Robert St.Pierre and Tracy Rimdzius. *A Study of Classroom Literacy Interventions and Outcomes in Even Start*. National Center on Education and the Economy. 2008. URL: `https://ies.ed.gov/ncee/pubs/20084028/pdf/20084028.pdf`.

[3] David Dickinson. "Speaking Out for Language: Why Language Is Central to Reading Development". In: *American Educational Research Association* 39 (2010).

[4] Alexandre Gramfort Vincent Michel Bertrand Thirion Olivier Grisel Mathieu Blondel Peter Prettenhofer Ron Weiss Vincent Dubourg Jake Vanderplas Alexandre Passos Fabian Pedregosa Gael Varoquaux and David

Cournapeau. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[5] C. Cybele Raver Fuhua Zhai and Christine Li-Grining. “Classroom-based Interventions and Teachers’ Perceived Job Stressors and Confidence: Evidence from a Randomized Trial in Head Start Settings”. In: *Early Childhood Research Quarterly* (2012).

[6] Kosuke Imai and Michael Lingzhi Li. *Experimental Evaluation of Individualized Treatment Rules*. 2019. eprint: arXiv:1905.05389.

[7] Kosuke Imai and Marc Ratkovic. “Estimating treatment effect heterogeneity in randomized program evaluation”. In: (2013). DOI: 10.1214/12-AOAS593. eprint: arXiv:1305.5682.

[8] Yuekai Sun Jason D. Lee Dennis L. Sun and Jonathan E. Taylor. “Exact post-selection inference, with application to the lasso”. In: (2013). DOI: 10.1214/15-AOS1371. eprint: arXiv:1311.6238.

[9] Barbara D. Goodson Jean Layzer Carolyn J. Layzer and Cristofer Price. *Evaluation of Child Care Subsidy Strategies: Findings from Project Upgrade in Miami-Dade County*. For the U.S. Dept. of Health and Human Services. 2007. URL: https://www.acf.hhs.gov/sites/default/files/opre/upgrade_miami_dade.pdf.

[10] James Kim and David Quinn. “The Effects of Summer Reading on Low-Income Children’s Literacy Achievement From Kindergarten to Grade 8: A Meta-Analysis of Classroom and Home Interventions”. In: *Review of Educational Research* (2013).

[11] J. Zhu L. Gunter and S. Murphy. “Variable Selection for Qualitative Interactions”. In: (2011). DOI: 10.1016/j.stamet.2009.05.003.

[12] Jean Layzer. *Project Upgrade in Miami-Dade County, Florida, 2003-2009*. Inter-university Consortium for Political and Social Research (distributor). 2011. DOI: 10.3886. URL: https://doi.org/10.3886/ICPSR31061.v2.

[13] Joshua R. Loftus. *Selective inference after cross-validation*. 2015. eprint: arXiv:1511.08866.

[14] Christopher J. Lonigan and Timothy Shanahan. “Developing Early Literacy”. In: *National Institute for Literacy* (2009).

[15] Stephanie Al Otaiba and Barbara Foorman. “Early Literacy Instruction and Intervention”. In: *Community Lit J.* (2014).

[16] “Reading instruction grouping for students with reading difficulties”. In: ().

[17] Snyder Richard. "Resistance to Change among Veteran Teachers: Providing Voice for More Effective Engagement". In: *Journal of Educational Leadership Preparation,* 12 (2017). URL: https://files.eric.ed.gov/fulltext/EJ1145464.pdf.

[18] Skipper Seabold and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python". In: *9th Python in Science Conference*. 2010.

[19] Sebastian Suggate. "A Meta-Analysis of the Long-Term Effects of Phonemic Awareness, Phonics, Fluency, and Reading Comprehension Interventions". In: *Journal of Learning Disabilities* (2014).

[20] Jessica Vick Whittaker Tamara Halle and Rachel Anderson. *Quality in Early Childhood Care and Education Settings: A Compendium of Measures, Second Edition*. US Department of Health and Human Services. 2010. URL: https://www.acf.hhs.gov/sites/default/files/opre/complete_compendium_full.pdf.

[21] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: (1996).

[22] Stefan Wager and Susan Athey. *Estimation and Inference of Heterogeneous Treatment Effects using Random Forests*. 2015. eprint: arXiv:1510.04342.

8 Appendix

8.1 Appendix *A*.1: All Coefficients and Post-Selective Confidence Intervals

Note that the confidence intervals do not seem to match the standard deviation because conditional on the selection event, the correct confidence intervals are no longer symmetric (see [8]). All point estimates (effect sizes) are standardized. For continuous covariate, for example, an effect size of 0.5 would mean that a 1 standard deviation increase in the covariate is associated with a 0.5 standard deviation increase in the outcome. (This association is causal if applied to a treatment/interaction term.)

8.2 Appendix *A*.2: Residual Analysis

Table 3: Literacy Resources Response Coefficients (Post-Selective)

Variable	pval	point	lower	upper	sd
Teacher Edu (Some College) (interaction)	0.416	0.827	-1.229	1.611	0.412
Teacher Edu (Some College) (interaction)	0.293	1.019	-0.998	1.876	0.448
Teacher Edu (College) (interaction)	0.537	0.675	-1.396	1.411	0.390
Treatment	0.495	-0.633	-3.515	1.678	0.387
2003_literacy_resources	0.001	0.642	0.331	0.936	0.124
intercept	0.262	0.028	-0.319	3.120	0.142
interaction_2003_Arnett_PosPunDet	0.983	0.054	-0.477	0.254	0.105
Pct Spanish Speaking (interaction)	0.808	0.064	-0.349	0.343	0.109
interaction_2003_literacy_resources	0.023	-0.443	-0.771	-0.071	0.167

Table 4: Oral Language Response Coefficients (Post-Selective)

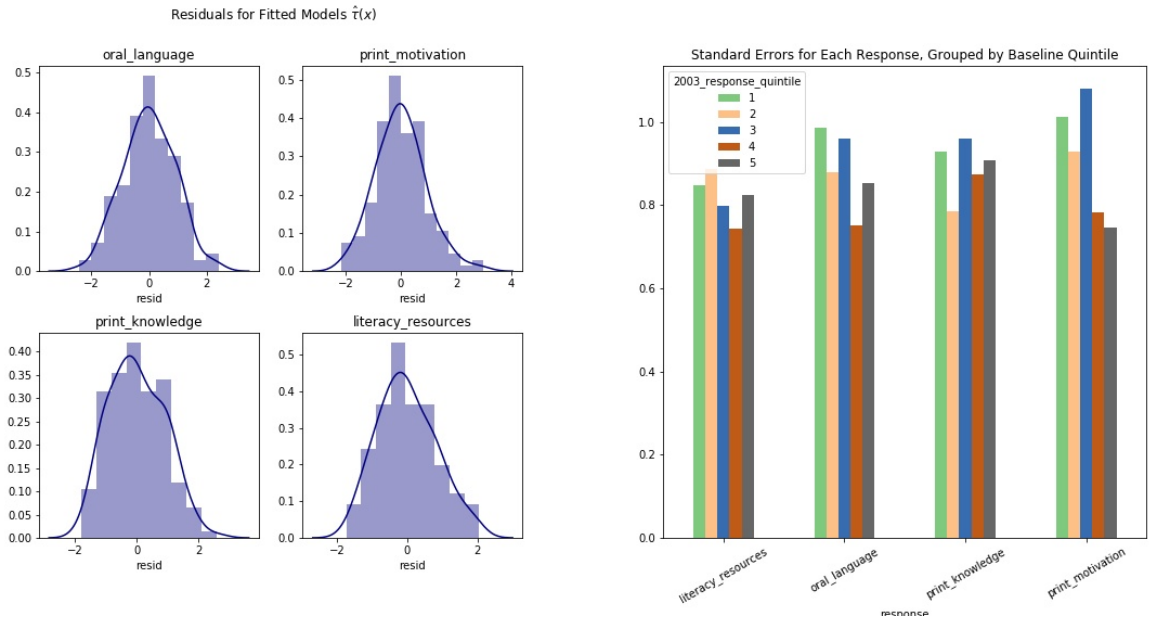
Variable	pval	point	lower	upper	sd
Teacher Edu (Some College) (interaction)	0.460	0.177	-11.608	2.751	0.318
Teacher Edu (College)	0.827	-0.111	-2.158	3.406	0.183
Treatment	0.810	0.596	-2.328	2.024	0.178
2003_Arnett_PosPunDet	0.117	0.235	-17.954	0.225	0.148
Pct Spanish Speaking	0.798	-0.047	-0.979	0.702	0.086
2003_oral_language	0.020	0.350	0.112	4.228	0.145
intercept	0.790	-0.337	-0.901	1.196	0.181
interaction_2003_Arnett_PosPunDet	0.072	-0.155	-0.119	26.354	0.179
interaction_2003_oral_language	0.028	-0.393	-5.409	-0.094	0.175

Table 5: Print Motivation Coefficients (Post-Selective)

Variable	pval	point	lower	upper	sd
Teacher Edu (Some College) (interaction)	0.751	0.085	-1.599	2.594	0.311
Teacher Edu (Some College) (interaction)	0.517	-0.247	-1.142	5.486	0.361
Teacher Edu (College)	0.972	-0.243	-1.031	1.711	0.227
Treatment	0.328	0.599	-0.656	1.134	0.196
2003_print_motivation	0.007	0.377	0.116	0.618	0.123
intercept	0.400	-0.255	-1.485	0.524	0.198
interaction_2003_Arnett_PosPunDet	0.500	0.119	-0.242	0.345	0.100
Pct Spanish Speaking (interaction)	0.387	0.021	-1.659	0.251	0.105
interaction_2003_print_motivation	0.011	-0.474	-0.794	-0.123	0.163

Table 6: Print Knowledge Coefficients (Post-Selective)

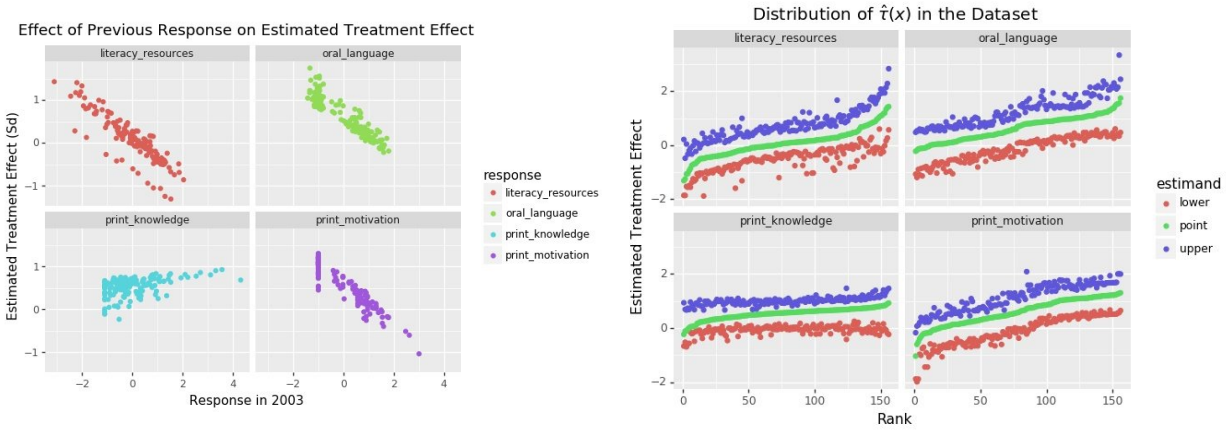
Variable	pval	point	lower	upper	sd
Teacher Edu (Some College)	0.114	-0.166	-0.282	19.320	0.241
Treatment	0.044	0.528	0.016	0.860	0.170
Pct Spanish Speaking	0.904	-0.151	-1.039	1.110	0.087
2003_print_knowledge	0.231	0.109	-0.343	3.209	0.141
intercept	0.022	-0.386	-2.821	-0.095	0.145
interaction_2003_Arnett_PosPunDet	0.862	0.193	-0.861	0.676	0.103
interaction_2003_print_knowledge	0.287	0.059	-4.715	0.457	0.173



(a) A subfigure

(b) A subfigure

Figure 2: Residual Analysis: Figure 2a plots the distribution of the residuals for each fitted model of $\hat{\tau}(x)$ to demonstrate they are approximately Gaussian. Figure 2b plots the empirical standard deviations of the residuals, grouped by the quintile of the pre-treatment baseline response level. These standard deviations are approximately equal, indicating that the homoskedacicity assumption is plausible.



(a) Estimated Treatment Effects versus Baseline OMLIT

(b) Estimated Treatment Effects

Figure 3: Visulaizing the ITR: Figure 2a plots the estimated treatment effect against the baseline OMLIT score from 2003. Figure 2b plots the estimated treatment effects $\hat{\tau}(x)$ of every center in the dataset. Note that in calculating PAPE and PAPD, we only use 60% of the data to fit $\hat{\tau}$, but for visualization purposes we graph all the centers.