Heterogenous Treatment Effects in Early Language Literacy Interventions

A Post-Selective Approach

Alex Chin and Asher Spector

November 14, 2019

Abstract (49 Words)

Early language literacy (ELL) interventions are known to improve educational outcomes on average [Dickinson, 2010]. However, policymakers need to know how different subgroups may be affected by ELL interventions. While researchers have recently developed tools to identify heterogenous treatment effects, most assume the availability of vast amounts of data. Thus, current literature analyzing heterogenous treatment effects for ELL interventions is sparse: previous studies rely on heuristic cross-study comparisons or ad-hoc subgroup analysis.

To fill this gap, we apply sparse LASSO techniques to data from ELL interventions to rigorously identify heterogeneous effects. We find that classrooms with initially lower support for language-learning gain more from these interventions. Our main results are cluster-robust, account for multiple-comparisons, and do not rely on asymptotic theory. This contributes to the literature by first rigorously demonstrating the equalizing effects of ELL interventions, and second demonstrating an effective application of sophisticated heterogenous methods on a small dataset.

Outline

1. Theory and Hypothesis
2. Contributions: Methodological and Empirical Literature Review
3. Data: Summary of Original Study
4. Methods: Sparse LASSO
5. Results: Equalizing Effects
6. Next Steps

Theory and Hypothesis

Early language literacy (ELL) interventions are teacher training and support programs which aim to increase literacy of young children (2-5 years).

Research Question:

1. Do ELL interventions have different causal effects on different subgroups of students and teachers?
2. Can we identify such subgroups in a principled, automatic fashion?

Findings: After analyzing randomized trials from Miami, we found

1. Interventions have stronger positive effects on classrooms with weaker support for language-learning
2. This interaction effect accounts for the majority of the treatment effect (we'll formalize this later)

Contributions: Gaps in Literacy Intervention Literature

- Most literature attempting to analyze heterogenous effects focus on meta-analyses that code studies and compare effect sizes.
  - This requires using heuristics to compare diverse outcome measure across studies.
  - The interventions also differ across studies, further complicating comparison.
- Many observational and experimental studies suffer from multiple testing issues when doing subgroup analysis [Layzer, 2011]
- “These findings suggest that identifying subgroups is important in developing and evaluating the effectiveness of reading comprehension interventions” [Kristen L. McMaster and Carlson, 2011]
- Other literature supports the importance of investigating heterogenous effects in reading interventions [Suggate, 2014]

Contributions

Broadly, to our knowledge, previous studies of heterogenous treatment effects for literacy interventions are sparse and largely rely on ad-hoc subgroup analyses or meta-analysis.

With this in mind, we hope that our research offers two main contributions:

- First, it more rigorously demonstrates that the treatment effects of language literacy interventions is highly heterogenous, with treatment effects being substantially more positive with weaker literacy support.
- Second, we hope that this application might demonstrate that more sophisticated methods causal inference methods (e.g. sparse regression/post-selective inference) can be effective in contexts without massive amounts of data. (In this case, $n = 165$).

Data: Overview

- Initial study randomly assigns classrooms in 162 child care centers for low income individuals to one of three ELL interventions or a control group (continuing normal reading curriculum) from Fall 2003 to Spring 2005.
  - Breakdown by Intervention: 38 Ready Set Leap, 36 BELL, 36 Breakthrough to Literacy, 55 Control.
  - We default to the domain knowledge of the original study's researchers who expected similar enough results from the three treatments to analyze them as a **singular treatment group**. We do the same in our analysis.
- The original study's goal was to measure the effect of these treatments on **classroom support for literacy development** and **students' language skills** (see **Data: Measures of Outcome** for the exact outcome variables).
- Individual and classroom-level covariates collected.

Data: Measures of Outcome

1. At the classroom-level, teacher and classroom **support for literacy development** (Spring ’04, and Spring ’05) is measured using the standard test Observation Measures of Language and Literacy Instruction (OMLIT). Six variables are analyzed as outcomes (but we use four since baseline data is missing for 2): **support for literacy resources, support for literacy activities, support for oral language, and support for phonological awareness**.
2. At the student-level, student’s **language skills** and pre-literacy indicators (Spring ’05) are measured using the widely-used Test of Preschool Emergent Literacy (TOPEL). This gives an **Early Literacy Score** for each student.

Note that up to this point we have only analyzed the first outcome of support for literacy development. We are working on cluster-robust methods to analyze student-level outcomes (see Next Steps).

Data: Select Covariates

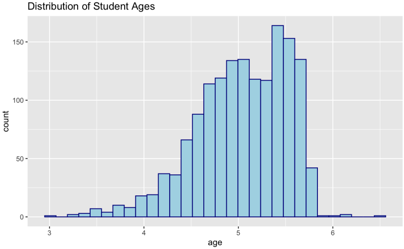
A number of covariates were recorded both at the individual and classroom levels. As such, the lists below give larger categories, each of which may describe multiple of the measured covariates.

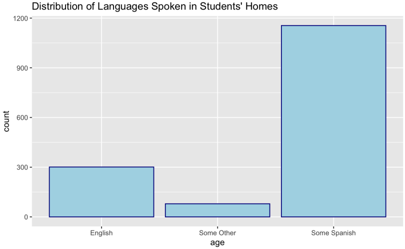
Covariates

1. Student Level: Age, sex, and language spoken at home.
2. Teacher (classroom) Level: Teacher preferred language, teacher education level, percentage of students speaking various languages, LAP-D (analyzes young child skills from motor control to social-emotional ability) scores (Fall ’04), OMLIT baseline scores (Fall ’03), size of center, Arnett Caregiver Rating Scale (Arnett) measured caregiver’s emotional support of students.

Student Level Covariate Distributions

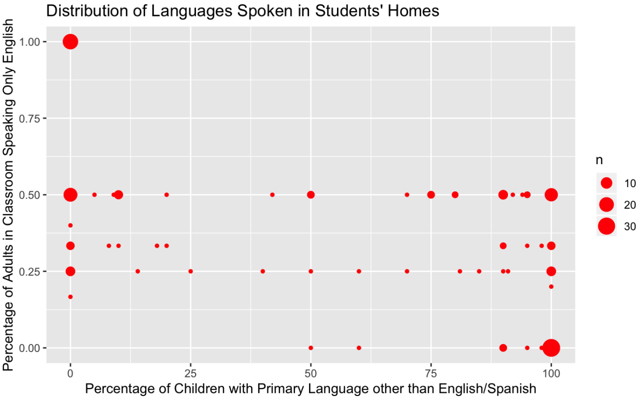
- About 50/50 Male Female ratio.
- High numbers of households speaking some Spanish.
- Overall expected distributions of other covariates (age plotted below as an example).


Distribution of Student Ages


Distribution of Languages Spoken in Students' Homes

Classroom Level Trends

- General trend of more diversity in students' languages associated with more diversity in teachers' languages.
- This is mainly driven by classrooms with all-multilingual or all-English-speaking students and teachers.



Methods: Setup for Heterogenous Treatment Effects

[Imai and Ratkovic, 2013]

Where $Y_i \in \mathbb{R}$ is the response, $X_i \in \mathbb{R}^p$ are pre-treatment covariates, and T_i is the treatment, our estimand is $\tau(x)$:

$$\tau(x) = E[Y_i(1)|X_i = x_i] - E[Y(0)|X_i = x_i]$$

In practice, this may be too hard to estimate with many covariates (see [L. Gunter and Murphy, 2011]).

Instead, we aim to identify a sparse subset of covariates $\tilde{X}$ and estimate

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

following [Imai and Ratkovic, 2013]. Note as a result, *which estimand we are analyzing* is random and selected by the procedure. This is common in selective inference ([Lee et al., 2013]).

Identification Assumptions

We make the following standard identification assumptions:

1. There is no interference between clusters
2. Each cluster has a nonzero probability of assignment to treatment/control
3. The treatment is independent of potential outcomes

Note that:

1. Condition 1 is likely to hold because the clusters are phsyically separate education centers
2. Conditions 2 and 3 are guarenteed to hold because we analyze data from a randomized experiment.

Methods: Estimation Set-Up

This largely (but not entirely) follows [Imai and Ratkovic, 2013]

For each response at the cluster (teacher) level, denote the set of pre-treatment covariates as X_i.

- Note for this application, X_i includes pre-treatment response levels, the teacher education level, and student demographics (around 40 features).
- For our clustered data ($n = 160$), this is too many features

For each covariate X_{ij}, we include the interaction term $Z_{ij} = T_i \cdot X_{ij}$, as well as the binary treatment, and model the response

$$Y_i = \mu + \beta^T Z_i + \gamma^T X_i + \epsilon_i$$

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.

Methods: Objective Function

This largely (but not entirely) follows [Imai and Ratkovic, 2013]

We rely on the LASSO regression ([Tibshirani, 1996]), which selects a sparse subset $\tilde{X}_i$ using the objective function:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \lambda_1 \sum_{j=1}^{p} \left|\hat{\beta}_j\right| + \lambda_2 \sum_{j=1}^{p} \left|\hat{\gamma}_j\right|$$

The l_1 penalizations ensure a sparse solution, so the Lasso selects our $\tilde{X}_i$.

We tune our values λ_1 and λ_2 using cross-validation by performing a grid-search.

1. Note unlike [Imai and Ratkovic, 2013], our CV statistic is simply the MSE, and does not reward sparsity. (This will be crucial for obtaining p-values later).

Methods: Inference Goals

After running the LASSO, we would like to do two types of inference:

1. Given nonzero treatment interaction coefficients $\hat{\beta}_i$ for $Z_{ij} = T_i \cdot X_{ij}$, we'd like to calculate (exact) confidence intervals for $\hat{\beta}_i$.
   (a) We can think of each $\hat{\beta}_i$ as a directly interpretable component of $\tau(\tilde{x})$.
   (b) We use the post-selective inference approach of [Lee et al., 2013] for this.

2. Additionally, we'd like to calculate confidence intervals for $\tau(\tilde{x})$, our main parameter of interest.
   (a) We use the bootstrap to do this.

The next two slides elaborate on these two methods (respectively).

Methods: Exact *p* values and Post-Selective Inference

After selecting an active set of variables $\mathcal{B} = \{j : \hat{\beta}_j \neq 0\}$ and $\mathcal{G} = \{j : \hat{\gamma}_j \neq 0\}$, regress Y_i on $Z_{i,\mathcal{B}}, X_{i,\mathcal{G}}$ to obtain post-selective coefficients $\hat{\beta}_S$ and $\hat{\gamma}_S$.

This is (almost) the cardinal sin of statistics: how do we get valid *p*-values for $\hat{\beta}_S$ and $\hat{\gamma}_S$, since selected variables will have artificially low *p*-values?

[Lee et al., 2013] define an adjustment method by which we can obtain exact uniform *p*-values for $\hat{\beta}_S$ and $\hat{\gamma}_S$ in finite samples

- Their methods are beyond the scope of this presentation, but briefly they characterize the LASSO selection event as an affine constraint on Y which induces a truncated Gaussian distribution.
- [Loftus, 2015] extend this to the case where the LASSO parameters are selected via cross-validation.

Methods: Estimating $\tau(\tilde{X})$

Under the selected model, we see that

$$E[Y_i(1) - Y_i(0)|\tilde{X}_i] = \beta_S \tilde{X}_i$$

Since $\hat{\beta}_S$ will converge to β_S (just by the properties of linear regression), we can estimate

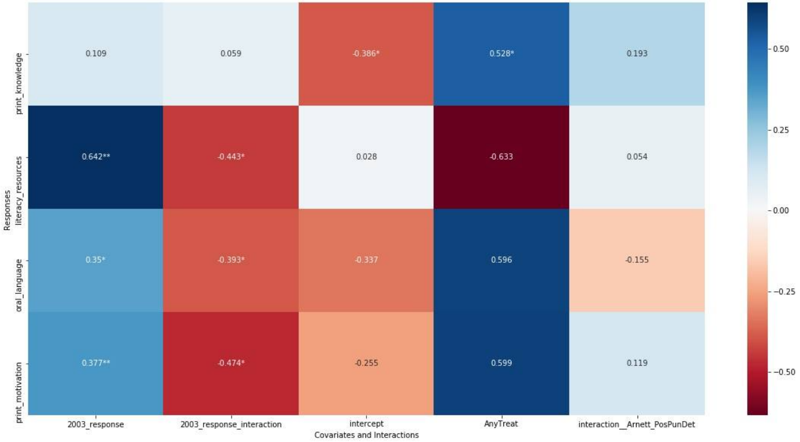
$$\hat{\tau}(\tilde{x}) = \hat{\beta}_S^T \tilde{x}$$

To conduct large-sample confidence intervals, we can use the bootstrap. (Unfortunately, this is rather computationally intensive.)

To be clear: these bootstrapped confidence intervals are *not* post-selective because the bootstrapped procedure involves re-selecting the model. As a result, they have full marginal coverage guarentees.

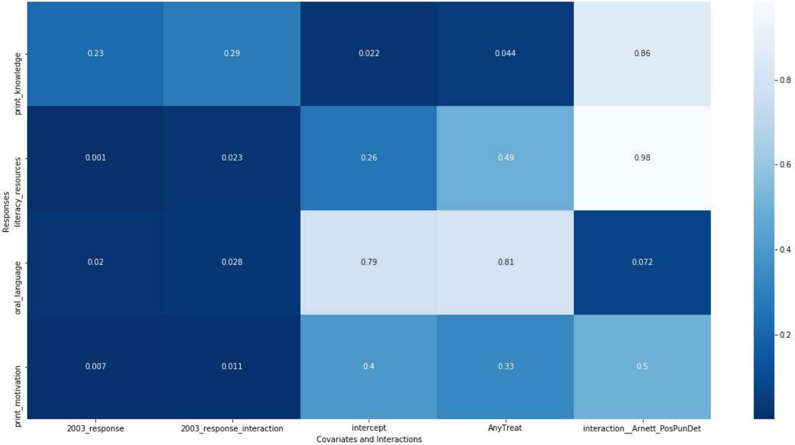
Results Part 1: Interaction Terms

In the next two slides, we plot the (post-selective) *p*-values and coefficients for all four outcome variables. We plot only the terms that were selected for all outcome variables (which include all statistically significant terms). * indicates a statistically significant p-value.



Results Part 1: Interaction Terms

Below are the p-values corresponding to the above values. Note that for 3/4 of the outcomes, both baseline covariates and interactions between baseline and treatment values are statistically significant, even accounting for multiple testing.

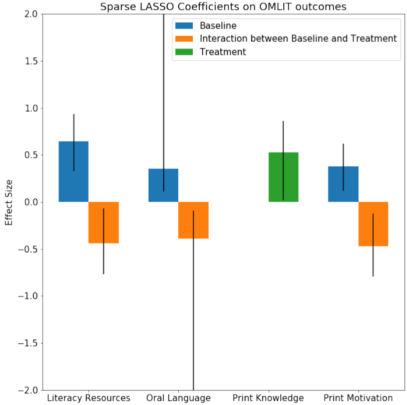


Responses	2003_response	2003_response_interaction	intercept	AnyTreat	interaction__Arnett_PosPunDet
print_knowledge	0.23	0.29	0.022	0.044	0.86
literacy_resources	0.001	0.023	0.26	0.49	0.98
oral_language	0.02	0.028	0.79	0.81	0.072
print_motivation	0.007	0.011	0.4	0.33	0.5

Results: Sparse LASSO Significant Coefficients

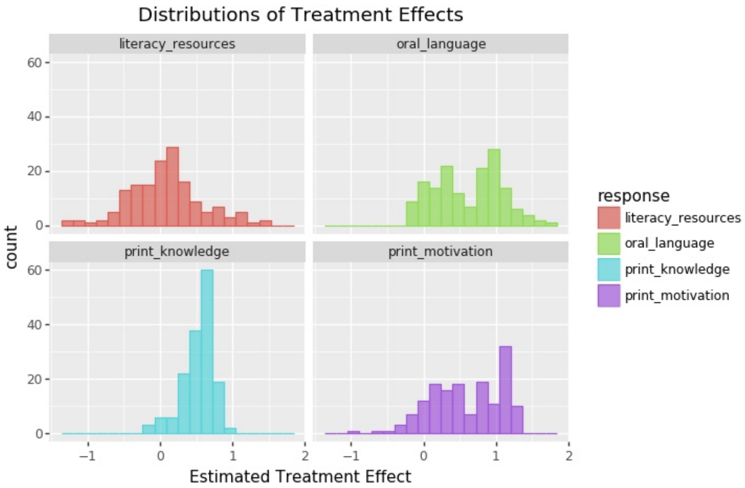
The below graph removes all not statistically significant coefficients to give a better picture.

- The 95% are such that we can be confident in our effects' directions.
- Note that in general, a larger baseline score indicates a larger post-treatment score.
- However, a large baseline score decreases the effect of treatment. This implies that the ELL intervention is equalizing.



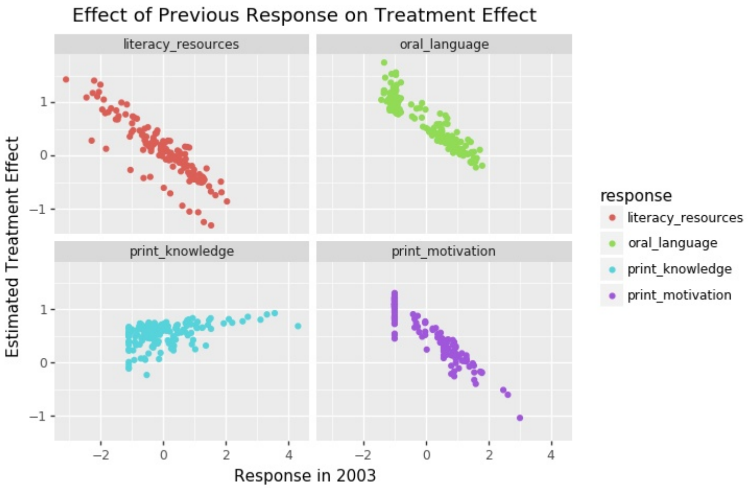
Results Part 2: $\hat{\tau}(\tilde{x})$

Below we present some summaries of the distributiones of $\hat{\tau}(\tilde{x})$, computed for the whole dataset. Note that the distributions are fairly wide...



Results Part 2: $\hat{\tau}(\tilde{x})$

...and as expected, are strongly correlated with the initial level of the response in 2003.



Summary

In summary:

1. Using data from a cluster-randomized experiment in Miami-Dade from 2003-2005, we find evidence that early language literacy interventions are drastically more effective in classrooms with lower support for language literacy.
2. Our results are exact, robust to model mispecification, and robust to multiple testing problems.
3. As far as we are aware, our results more rigorously demonstrate these effects than the previous literature, which relied largely on ad-hoc subgroup analysis and meta-analyses.

Next Steps

- The individual-level outcomes present the difficulty of cluster randomization (not the original study did not have cluster-robust standard errors in their analysis, but we want our heterogenous analysis to be robust with respect to clusters). We will continue to try and find a method that allows us to estimate heterogenous treatment effects in a cluster-robust manner.
- Our current methods have the advantages that they yield exact *p*-values, are robust to nonlinear responses, and account for multiple testing problems. However, they're not heteroskedacitity robust, so it would be nice to look into that.
- We also should probably find better ways to graphically display our results :) We are in particular open to suggestions about how to display the $\hat{\tau}(\tilde{x})$ values.

Citations I

[Dickinson, 2010] Dickinson, D. (2010).
Speaking out for language: Why language is central to reading development.
American Educational Research Association, 39.

[Imai and Ratkovic, 2013] Imai, K. and Ratkovic, M. (2013).
Estimating treatment effect heterogeneity in randomized program evaluation.

[Kristen L. McMaster and Carlson, 2011] Kristen L. McMaster, Paul van den Broek, C. E. M. J. W. D. N. R. P. K. C. M. B.-G. and Carlson, S. (2011).
Making the right connections: Differential effects of reading intervention for subgroups of comprehenders.

[L. Gunter and Murphy, 2011] L. Gunter, J. Z. and Murphy, S. (2011).
Variable selection for qualitative interactions.

Citations II

[Layzer, 2011] Layzer, J. (2011).
Project upgrade in miami-dade county, florida, 2003-2009.
https://doi.org/10.3886/ICPSR31061.v2.

[Lee et al., 2013] Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2013).
Exact post-selection inference, with application to the lasso.

[Loftus, 2015] Loftus, J. R. (2015).
Selective inference after cross-validation.

[Suggate, 2014] Suggate, S. (2014).
A meta-analysis of the long-term effects of phonemic awareness, phonics, fluency, and reading comprehension interventions.

[Tibshirani, 1996] Tibshirani, R. (1996).
Regression shrinkage and selection via the lasso.