

Scale Up of an Influential Early Childhood Education Program*

Andrés Hojman[†]
Pontificia Universidad
Católica de Chile

Chanwool Kim
University of Chicago

Juan Pantano
University of Hong Kong

September 10, 2026

Abstract

The Early Head Start (EHS) program covers almost 200,000 children aged 0–2 in the U.S. It was built on many of the same principles as the highly successful Abecedarian (ABC) program. However, results from EHS’s early evaluation were considered weak, leading to claims that this was another example of a program that works well at the pilot stage but disappoints when implemented at scale. We demonstrate that this conclusion was perhaps premature. When comparisons are made appropriately, EHS effect sizes are as strong as those from ABC, if not larger. To compare the programs, we use forest-based methods to adjust for population heterogeneity and account for differential non-compliance.

*We thank the seminar participants at the University of Arizona, University of Chicago, Singapore Management University, Pontificia Universidad Católica de Chile, Indiana University, and KAIST.

[†]Pontificia Universidad Católica de Chile, Centro de investigación e innovación en niñez, adolescencia, resiliencia y adversidad (IINARA), and Centro para el bienestar y desarrollo de la adolescencia y niñez en la era digital (BAND). Thanks to ANID for its support through IINARA (ANID CIN 250068) and Fondecyt 1262318.

1 Introduction

A growing literature in economics emphasizes the importance of early childhood ([Elango et al. \(2015\)](#), [Currie and Almond \(2011\)](#)). More recently, the focus has shifted to very early childhood, when children are between the ages of 0 and 3. A particularly influential result that sparked interest in the 0-3 range came from evaluating the legendary Abecedarian (ABC) program ([Ramey \(1974\)](#), [García et al. \(2020\)](#)). At the frontier of this literature is the question of how to design early childhood interventions that succeed when implemented at scale ([List et al. \(2021\)](#)).

While it took three more decades than for preschool children ages 3 and 5, United States federal policy toward children ages 0-3 gained initial momentum in the mid-1990s with the launch of the Early Head Start (EHS) program. EHS was inspired by and modeled after the Abecedarian program. The program started in 1995 with 68 programs and grew to more than 700 by 2004. Researchers selected a few of the 68 initially funded programs as sites for a randomized evaluation of EHS in 1996. The EHS evaluation results were presented in a series of articles, most prominently in [Love et al. \(2005\)](#). The results were judged disappointing because the reported effect sizes did not approach the magnitudes obtained two decades earlier in the evaluation of ABC. Skeptics of the value of early childhood education programs quickly leveraged these lukewarm findings to launch strong criticisms of the program ([Besharov et al. \(2011\)](#), [Armor \(2014\)](#)), calling for a halt in its expansion or even its defunding. Some of these evaluation concerns may have helped stifle plans for EHS expansion. In the end, EHS never reached the scale of the Head Start program, which targets 3- and 5-year-olds and is modeled after the Perry Preschool Project. The concerns about the scalability of EHS echoed similar concerns with other social programs that appear to work quite well in pilot or demonstration settings but do not deliver the same impacts when deployed at scale, as argued in [List \(2021\)](#), [Moffitt \(2006\)](#) and in [List et al. \(2021\)](#); [Al-Ubaydli et al. \(2021\)](#) for early childhood programs, in particular. The EHS evaluation team responded to and rebutted some of the attacks, but in the end, the sense that EHS was not worthy of expansion became entrenched in policy circles.

In this paper, we revisit the EHS evaluation, focusing on a comparison of its effectiveness relative to ABC. We use the original microdata from both evaluations to replicate earlier results, then step through a series of adjustments to make what we argue are “apples-to-apples” comparisons between the programs to judge their relative effectiveness. We first focus on the subset of EHS sites that implemented programs more similar to ABC’s; we then work on a subsample of EHS that more closely resembles ABC’s. We then move beyond the previous ITT analyses to account for differential non-compliance by estimating the Local

Average Treatment Effect. In both the ITT and LATE cases, we also use the machine-learning causal-inference techniques proposed in [Athey et al. \(2019\)](#) to model treatment-effect heterogeneity using random forests and estimate these parameters. The forests then help us identify which observables drive treatment effect heterogeneity for EHS. We then use these forests to answer an interesting “what if” that improves comparability across estimates for both programs. We ask: what if EHS were implemented in a population that looks like ABC on those key drivers of heterogeneity? The forests allow us to re-estimate ITT and LATE under that counterfactual distribution of key observable heterogeneity drivers.

We then argue that even a LATE approach that adjusts for differential non-compliance across programs commingles two treatment-effect parameters, obscuring what is perhaps the ultimate comparison of interest. We argue that this “overall” effect combines the effect of EHS relative to keeping the child at home, and the one relative to sending the child to alternative centers. We refer to these two distinct average treatment effects for the corresponding subset of compliers as “subLATEs.” While these are not point-identified, we show how we can estimate bounds for them under reasonable assumptions, providing additional insight into the sources of the gap between ABC and EHS.

By carefully comparing the pilot (ABC) and initial scaled-up (EHS) version of a program, our work contributes to an emerging literature popularized in [List \(2021\)](#) that studies why promising interventions do not appear to work as well when deployed at scale. As discussed in [List et al. \(2021\)](#), this is particularly relevant in our context, early childhood interventions, where oftentimes programs show great promise in the pilot stage but appear to lose voltage when deployed more widely by policy-makers. Our work carefully unpacks the role of different populations, different treatment effect estimands, and even different fall-back options in the absence of treatment when assessing the apparent voltage drop reported when comparing Abecedarian and EHS. By rigorously re-evaluating this initial scale-up, our work complements work that explores issues associated with further expansion of this key federal intervention in early childhood ([Lim Brodowski and Naom \(2021\)](#)).

Our central goal in the main analysis is to explore the *vertical* generalizability, in the sense of [List \(2026\)](#), of an early childhood program like ABC as it scaled up into a more massive implementation like EHS. That is, we are focused on asking: is the program less, equally, or more effective as it scales up and reaches more people *in the same population*? We want to know whether the program scaled up effectively when evaluated on a comparable population but at a larger scale, i.e., dozens of centers instead of just one. While we do as much as we can to homogenize the population, EHS was implemented across a much wider geography and twenty years later. We cannot hold time, context, environment, and geography constant. However, we can adjust for the differences in program participation

generated by these factors. We implement LATE, sub-LATE, and Counterfactual LATE analyses to account for contextual factors that affected compliance, including differences in the availability of alternative childcare centers, the broader geographic scope of EHS, and cohort changes in maternal labor force participation.

The rest of the paper proceeds as follows. Section 2 discusses the two programs we examine: Abecedarian (ABC) and Early Head Start (EHS). We also describe the data we use from each. Section 3 steps through a series of adjustments that provide successively more appropriate ways to judge the effectiveness of EHS vis-a-vis that of ABC. Section 4 presents the more novel methodological discussion on the subLATEs and further adjusts for differences in the relative sizes of the two complier subgroups associated with the subLATEs in ABC and EHS. Section 5 discusses the external validity of our results, and Section 6 concludes.

2 The Data

The data in this paper come from two early childhood education programs: (i) ABC and (ii) EHS.¹

The Carolina Abecedarian Project (ABC) The Carolina Abecedarian Project (ABC) targeted disadvantaged, predominantly African-American children in Chapel Hill/Durham, North Carolina. The program sought to enhance the early life skills of disadvantaged children. It promoted language, motor, and cognitive development as well as socio-emotional competencies, which were considered crucial for school success, including task orientation, ability to communicate, independence, and pro-social behavior (Sparling (1974); Ramey (1974); Ramey et al. (1985); Wasik et al. (1990)).

Early Head Start. Early Head Start (EHS) is a federally funded program for low-income families with children up to age 3 in the United States. ABC and its research findings explicitly inspired EHS. Center-based EHS can be viewed as a large-scale implementation of many of the principles established by ABC, with close similarities in age of onset, developmental and curricular approach, child-staff ratios, individualization, and the provision of intensive center-based care.

The national Early Head Start Research and Evaluation study began in 1995, around the same time the first Early Head Start programs were funded. It was designed to provide a

¹The analyses in this paper were not pre-registered. Because they rely on previously conducted experiments, prospective pre-registration before treatment assignment was not possible. Our main analytical choices nevertheless follow a principled structure: the EHS subsample approximates the ABC population, the causal forest procedure selects the continuous dimensions of heterogeneity, and the main estimands are standard ITT and LATE parameters.

rigorous, comprehensive evaluation of the impacts of Early Head Start programs. Seventeen (17) Early Head Start programs (sites) participated in the study. The programs were classified as (1) home-based, (2) center-based, and (3) combination (in which families can receive both home visits and center experiences or just one of the two). Although poverty-level families with children up to age 3 are eligible for Early Head Start, only low-income families with children up to 12 months old at enrollment were eligible for the evaluation. The programs independently determined through their application process that families met the Early Head Start eligibility guidelines. Then, for each site, they randomly assigned families to the program or the control group (with equal probabilities and no within-site stratification). Control group families were not allowed to receive Early Head Start services until their applicant child turned 3. They could, however, receive any other services available in their community. Sample enrollment and random assignment started in July 1996 through September 1998. Overall, 2977 families were randomly assigned to the program (1,474) and control (1,503) groups. The samples include between 140 and 222 families per site.²

Table 1 summarizes and compares the main characteristics of the two programs. First, while ABC was implemented in the mid-1970s, EHS was implemented in the late 1990s. This matters, particularly because the availability of early care and education options increased substantially over this period, which we account for below in our analysis. Second, ABC was implemented only in Chapel Hill, North Carolina, whereas the EHS evaluation was more geographically widespread and implemented across 17 sites. Third, while ABC was exclusively center-based, EHS included center-based, home-based, and mixed or hybrid programs. Finally, the adult-child ratio in ABC varied from 1-to-3 for infants to 1-to-4 for 2- and 3-year-olds. In EHS, it was 1-to-4 for all ages in the 0-to-3 range.

²Description from the Early Head Start Research and Evaluation study: User Guide ([United States Department Of Health And Human Services. Administration For Children And Families, 2004](#)).

Table 1: Characteristics of the Programs and Participants

Program	ABC	EHS		
Programme Overview				
Years Implemented	1972-1982	1996-2010		
Site	Chapel Hill, North Carolina	17 sites in the US		
Children (T:C)	111 (57:54)	2,977 (1,503:1,474)		
Programme Eligibility				
	HRI ≤ 11	Families with a pregnant		
	Biologically healthy	woman or a child under 12		
	No signs of mental retardation	months		
Treatment				
		Center- based programs	Home- based programs	Mixed programs*
Number of sites	1	4	7	6
Home Visits	No	No	Yes	Yes
Center Care	Yes	Yes	No	Yes
Center weeks per year	50	50		50
Programme Quality				
Adult-Child Ratio	1:3 (age 0-1)	1:4		
	1:4-5 (age 1-4)			
	1:5-6 (age 4-5)			
Teacher Certifications	HS grads, mixed	CDA specialising in infants and toddlers		

Notes: Years implemented include the total time children spent in the program. T:C represents the number of treatment and control children. HRI corresponds to the High-Risk Index; some items were changed in ABC. Source: extracted from [Elango et al. \(2015\)](#), [García et al. \(2020\)](#) and [United States Department Of Health And Human Services. Administration For Children And Families \(2004\)](#). * The mixed programs could have both visits and center-based care for all individuals or could give individuals a choice between the delivery modes.

2.1 Baseline Characteristics of the Populations

Table 2 describes the baseline characteristics of the populations served by the two programs we study. It also summarizes measures of the child’s IQ, the outcome variable we focus on (see the supplemental Appendix for details on the cognitive ability measures, covariates, and sample selection process). The table includes 8 columns: 6 for EHS and 2 for ABC. For EHS, we compute descriptive statistics. first for the full sample (“All”), then for the sample

of children in the “center only” and “hybrid” sites (“Center+Mixed”), and finally for the smaller sample of “Center Only” sites, which is most comparable to ABC. In each of these 3 EHS samples and the ABC sample, we further distinguish two columns. One is based on all observations denoted by “Full” and another denoted “Subsample” where we focus only on black children whose mothers have at most a complete high school education. This makes a big difference for the various EHS samples, but not much for ABC (observations only change from 103 to 98). As the table shows, EHS and ABC differ substantially in their populations. While EHS still had an over-representation of blacks compared to the overall population, ABC was almost entirely composed of blacks. Also, ABC had younger and less educated mothers. While the maternal IQ was much higher in the full EHS samples, once we restrict children to the more comparable subsamples, maternal IQ is more similar between EHS and ABC. Finally, the share of children who participated in alternative care arrangements is much higher in the EHS samples than in ABC, consistent with the growth of the early care and education industry in the two decades separating the two evaluations.

Table 2: Baseline Child and Household Characteristics Across ECE Programs

Program	EHS						ABC	
Type	All		Center + Mixed		Center Only			
Sample	Full	Subsample	Full	Subsample	Full	Subsample	Full	Subsample
Outcome								
IQ	83.1	78.7	81.6	75.7	80.9	74.9	92.2	91.5
Assignment and Participation								
% Randomized	53.0	52.3	52.7	52.2	58.4	60.0	51.5	51.0
% Participated (Any)	54.0	52.3	55.7	53.1	58.7	55.5	51.5	51.0
% Participated (Center)	16.1	15.8	26.6	22.4	44.0	37.3	51.5	51.0
% Alternative Care	41.8	49.1	42.7	44.9	35.5	35.5	16.5	15.3
Children’s Characteristics								
% Male	49.0	48.9	49.5	49.4	49.1	49.1	47.6	46.9
% Black	37.4	100.0	43.3	100.0	40.9	100.0	98.1	100.0
# Siblings	1.0	1.1	1.0	1.2	0.8	0.6	0.6	0.6
Gestational Age (Weeks)	39.3	39.4	39.3	39.4	39.4	39.6	39.5	39.4
Mother’s Characteristics								
Mother’s IQ	90.6	85.2	89.3	84.1	88.8	83.6	84.0	83.4
Mother’s Age	22.1	20.3	22.1	20.1	22.5	21.5	19.8	19.6
% HS Completed	31.3	38.5	29.8	33.9	30.9	41.8	29.1	30.6
% College Completed	27.8	0.0	28.0	0.0	28.8	0.0	3.9	0.0
% Father Figure at Home	21.6	6.0	18.8	6.1	18.7	7.3	27.2	26.5
% Above Poverty	11.4	7.5	12.1	9.0	12.0	10.0	11.9	10.4
Sample Size								
# Observations	1,267	348	768	245	375	110	103	98

Notes: ABC and EHS mothers’ IQ tests were based on the Wechsler Adult Intelligence Scale, measured at the time children were entering the program. Mothers’ years of education are measured at age 0 for ABC. For EHS, they are measured at age 2.

2.2 Balance

Both ABC and EHS were assigned experimentally, and we leverage these randomized designs to estimate program impacts. Appendix Tables A.2 and A.3 show that treatment and control groups appear to be balanced in baseline characteristics for ABC and EHS, respectively. Since our analysis focuses on EHS subsamples—restricting both program types and the populations included—it is reassuring that Tables A.4 and A.5 also show balance between treatment and control groups within these restricted samples.

2.3 Participation

In both programs, we use program participation measures. In ABC, they are explicit in the data. We have measures of participation in the ABC program and in alternative programs. In EHS, we must construct the participation variables from different pieces of data available in the surveys. We first create a variable for the number of months participating in center-based care. Then, we assign those months to EHS if the child reports participating in a center-based EHS program in any variable.³ For our subLATE analyses, we need a measure of participation in alternative centers. This analysis further distinguishes what happens when a child does not take either the ABC or EHS treatment. It requires classifying the family as either keeping the child at home (n) or sending him/her to an alternative center (c). Therefore, we need an observable measure of participation in such alternative centers to make operational this classification. We construct it as the number of months children not participating in EHS attend center-based care. We only use binary versions of these variables (does the child participate for at least m months).

2.4 Attrition

In ABC, attrition was very low: fewer than 10% of participants are missing the relevant test-score outcomes (Appendix Table A.1). In contrast, attrition in EHS was substantial. Among the population of interest—children in center-based programs with characteristics comparable to ABC—approximately 33% are missing either the age-three IQ outcome or program-participation information. Moreover, attrition differs by treatment status, with response rates higher among treated than control children. Section 3.5 discusses this issue in detail and presents our approach to addressing it.

3 An Apples-to-Apples Comparison between ABC and EHS

In this section, we go through a series of adjustments that, we argue, provide a successively more appropriate contrast between estimates of treatment effects for EHS and ABC. We start with the ITT effects of EHS on cognition using the full EHS sample that includes all types of EHS programs: center-only, mixed programs, and visit-only programs. We then

³The EHSRE contains different measures of EHS participation. Program staff reports imply very little control-group crossover (0.7%) (Love et al., 2002, p. 7 and n. 3), whereas the child-care tracking data identify approximately 10% of control children in center-based sites as receiving care in an EHS center. We use the tracking measure because it reflects the research team’s identification of children’s actual care providers. The available documentation does not establish the source of the discrepancy, but using the parental-interview measure instead leaves the estimated first stage almost unchanged and does not substantively affect our conclusions.

explore what happens when we focus on the EHS sites that run programs more similar to ABC. We then exclude certain demographics that do not overlap between the two studies. We then account for differences in the density of relevant observable heterogeneity across the two programs, even within the overlap region. We also account for differential non-compliance.⁴

3.1 EHS Program Type

As described in Table 1, there are different variants of EHS. While ABC was exclusively center-based and did not include any home-visit component, some EHS programs were hybrid, mixing center-based programming with home visits, while others were entirely home-visit based. Therefore, the first step in our comparability analysis is to explore how EHS results vary as we progressively narrow our focus on programs most similar to ABC. Table 3 compiles all of our main results. Column 1 presents the ITT results including all EHS programs in the first row. The second column reports results excluding “visits-only” programs, which differ substantially from those used in ABC, and the third column focuses on EHS programs that were exclusively center-based, as in ABC, excluding not only “visits-only” but also “mixed” programs.

As we can see, the EHS treatment effect grows as we progressively focus on more similar programs. In the full sample, which pools all sites regardless of program type, the effect size is small and not statistically significant; excluding the “home-visits-only” sites increases the effect size from 0.08 to 0.13, and this larger effect is already significant at the 10% level. Moreover, restricting the sample to the sites most comparable to ABC, that is, the sites that implemented “Center Only” programs, further increases the effect size to 0.18. Given its better comparability with ABC, we focus the remaining analyses on the “Center Only” variant of EHS, but the first and second columns of the table still present our results for the larger samples that include “mixed” and “home-visit-only” sites. To improve our precision, we use Flexible Regression Adjustment (FRA, [List et al. \(2024\)](#)),⁵ which makes

⁴In general, in multi-site experiments it is necessary to include site dummies for the analyses. An exception is when all sites have the same percentage of treated and controlled children. This is roughly true in our case: across all EHS sites, the percentage of treated children is always between 52.04% and 48.03%. Because of this, and to keep our methods simpler, we do not include the site indicators in our main analyses.

⁵In our setting, we apply FRA to the EHS and ABC samples with random assignment R as the treatment, child IQ as the outcome, and the baseline covariates listed in Table 2 (maternal IQ, race, child sex, maternal age and education, siblings, gestational age, and poverty status) as X , using two folds. For the ITT, we fit m_w by OLS within each arm. For the LATE, we apply the same procedure to both the outcome and the participation indicator D (fitting m_w with a random forest) and form the Wald ratio of the two regression-adjusted ITTs, with standard errors obtained using the delta method applied to the joint variance of the influence functions. This yields the precision gains reported in the FRA rows of Tables 3 and 4 while leaving the point estimates essentially unchanged.

the Center+Mixed specification significant at the 5% level. Given this advantage, and because the remaining results do not change substantially under this approach, we focus on the rows including FRA from now on.

3.2 Population Heterogeneity, Race and Maternal Education

In addition to restricting to more similar programs, we must account for differences in the populations served. First, we recognize substantial differences in race and maternal education across the two programs. As shown in Table 2, the target populations in the two programs differed substantially. In particular, while ABC participants were almost entirely black, EHS had a large number of non-black children; in fact, a vast majority of its study population. Also, while EHS included mothers with quite different education levels, all mothers in ABC had attained at most a high school diploma.

As a result, a natural step in comparing the pilot (ABC) and the scaled-up implementation (EHS) is to use an EHS subsample that excludes non-Black and highly educated mothers, since ABC has essentially no children with those characteristics. If the effects for those children differ substantially, keeping them would hinder comparability. In what follows, we refer to this subsample of Black children with low-educated mothers as the “Subsample.” Table 3, Row 4, Column 3 presents the ITT results for this more comparable subsample of EHS children. The effect size more than doubles to 0.47 and is now statistically significant at the 1% level. In the ABC column, as expected, the effect size does not vary much between rows 2 and 4, changing only from 1.09 to 0.95, since the “Subsample” restriction has almost no bite in the ABC sample.

3.3 Different Population Densities of Heterogeneity-Relevant Observables

If different individuals accrue different effects from participation in ECE programs and the populations differ, then a natural question is: what would be the ITT effect of EHS if randomized into a population like ABC’s? Previous work has documented substantial heterogeneity in treatment effects for different early childhood education programs, and EHS is no exception. We use a principled approach to extract relevant heterogeneity in EHS treatment effects. To that end, we draw on recent insights from the literature that leverage machine learning techniques to study causal inference questions. Following [Athey et al. \(2019\)](#), we estimate a generalized causal forest that provides a vehicle to obtain data-driven heterogeneous treatment effects. In this context, the forest predicts treatment effects across dimensions of observable heterogeneity rather than outcomes.

The causal random forest averages causal trees estimated on different bootstrap samples.

Each causal tree partitions the observable heterogeneity space differently to capture the most essential heterogeneity in treatment effects across the observed covariates. The forest-based ITT estimate then aggregates these leaf-level, individually heterogeneous effects. We first estimate a causal forest with all the covariates listed in Table 2. This causal forest can be used to learn which covariates in Table 2 are most important for capturing treatment-effect heterogeneity in EHS. This analysis shows that the mother’s IQ and the mother’s age are the two covariates most important for explaining heterogeneity in treatment effects. We then re-estimate the causal forest using only these two covariates to build the trees. Forest-based estimates of ITT effects for EHS following this procedure are presented in the ITT Panel, Row 5, Column 3. The effect size is 0.52, slightly larger than the 0.47 we obtained for the “Subsample” using the simple experimental analysis without covariates.

The causal forest also provides a device to further probe into the EHS and ABC results. To improve comparability, we can ask what ITT effect we would estimate for EHS if the joint distribution of heterogeneity-relevant covariates (i.e., maternal age and maternal IQ) in EHS were the same as in ABC. Using the estimated causal forest from EHS, we can estimate what that “what if” ITT would look like if EHS were implemented in a population with the same joint density of mother’s IQ and mother’s age as ABC. In practice, this means taking maternal age and IQ from the ABC “Subsample” of Black and less-educated mothers and placing them into the appropriate leaves of the EHS forest, providing an imputed treatment effect for each observation while preserving their joint density. The results after aggregating those estimates into a single average effect are presented in the ITT Panel, Row 6, Column 3. Although ABC’s mothers appeared slightly younger, the results are similar, with an effect of 0.53 instead of 0.52.

3.4 Non-Compliance

As in many other RCTs, assignment to treatment does not guarantee take-up of, or participation in, the assigned treatment. This is the problem of non-compliance. A further refinement in the comparison between ABC and EHS must then account for the possible differences in the degree of non-compliance. While compliance in ABC was quite high, compliance in EHS was much lower. This raises the question of how EHS compares to ABC after accounting for the differential extent of non-compliance. That is, even after adjusting for population differences, comparing ITT effects between the programs may obscure the true differences in the effects of actually *participating* in each program. For example, participation in EHS could be as or more effective than participation in ABC, yet the lower take-up rate of the EHS offer implies a greater dilution of the true effects of participation in the intent-to-treat analysis.

To the best of our knowledge, most of the previous analyses of the EHS program have focused on ITT effects. While they focused on ITT results, [Love et al. \(2005\)](#) reported results from [Bloom \(1984\)](#)’s estimator to address non-compliance, since some families that received the EHS offer did not enroll their children in EHS. The Bloom estimator essentially inflates the ITT estimate by dividing it by the EHS participation rate of those who received an EHS offer. However, Bloom identifies the Average Treatment Effect on the Treated (ATET) only if there are no always-takers. In other words, ATET is only identified if there is no two-sided non-compliance. Two-sided non-compliance could be present if those offered EHS decide not to participate, and those not offered EHS nonetheless manage to participate. Indeed, we observe that some families randomized to receive no EHS offer still have children participating in EHS.

Therefore, we rely on Wald/2SLS estimators instead of Bloom and interpret the identified parameter as the Local Average Treatment Effect (LATE) of EHS participation versus no EHS participation for the subpopulation of children who are induced to participate in the EHS program by receiving the EHS offer, following [Imbens and Angrist \(1994\)](#). Results for the subsample are reported in the LATE panel of Table 3, Row 4, Column 3. The effect size almost triples, jumping from 0.47 in the ITT analysis to 1.44 using the LATE. This estimate uses an EHS participation variable equal to one if the child ever participated in EHS. We probe the robustness of alternative definitions of EHS participation in the next two rows by defining EHS participation as participation for at least 6 months or at least 12 months. The results are similar, ranging from 1.10 to 1.44. The non-compliance problem in ABC was relatively small, with a complier share of 87%, and so the change between ITT and LATE is small, going from 0.95 to 1.06 in the “Subsample.” After this adjustment, the results show that, if anything, EHS effects were stronger than ABC’s, in contrast with the prevailing narrative based on less careful comparisons.

Table 3: ITT and LATE Effects of EHS and ABC
Standard and Forest-Based Estimates

		ITT				
Program	EHS					ABC
Type	All	Center + Mixed	Center Only			
Full	0.08 (0.05)	0.13* (0.07)	0.18* (0.09)			1.05*** (0.18)
Full/FRA	0.09 (0.05)	0.13** (0.07)	0.17* (0.10)			1.09*** (0.20)
Subsample	0.13 (0.09)	0.22** (0.10)	0.52*** (0.14)			0.98*** (0.19)
Subsample/FRA	0.15 (0.09)	0.21** (0.10)	0.47** (0.19)			0.95*** (0.22)
Subsample/Causal Forest	0.14 (0.09)	0.23** (0.11)	0.52*** (0.15)			0.95*** (0.19)
Subsample/Causal Forest (ABC)	0.11 (0.10)	0.29*** (0.10)	0.53*** (0.13)			-
		LATE				
Program	EHS					ABC
Type	All	Center + Mixed	Center Only			
Participation	Any	Any	Any	6m	12m	12m
Full	0.28 (0.19)	0.29* (0.15)	0.32* (0.17)	0.31* (0.16)	0.32* (0.16)	1.19*** (0.20)
Full/FRA	0.33 (0.20)	0.26* (0.15)	0.27 (0.17)	0.25 (0.16)	0.26 (0.16)	1.15*** (0.20)
Subsample	0.55 (0.38)	0.64** (0.30)	1.51*** (0.51)	1.18*** (0.35)	1.18*** (0.35)	1.10*** (0.21)
Subsample/FRA	0.52 (0.40)	0.60** (0.30)	1.44*** (0.52)	1.10*** (0.35)	1.10*** (0.35)	1.06*** (0.22)
Subsample/Instrumental Forest	0.12 (0.37)	0.61** (0.27)	1.22*** (0.47)	1.24*** (0.35)	1.24*** (0.35)	1.15*** (0.22)
Subsample/Instrumental Forest (ABC)	0.25 (0.33)	0.64*** (0.23)	1.47*** (0.47)	1.33*** (0.33)	1.33*** (0.33)	-
Sample Size: Full	1,267	768	375			103
Sample Size: Subsample	348	245	110			98

Notes: Participation: “Any” refers to participating in any part of the program, whereas 6m and 12m refer to participating in at least 6 or 12 months of the program, respectively. The FRA rows use Flexible Regression Adjustment (List et al., 2024) to improve precision. The subsample excludes non-black children and highly educated mothers. Causal Forest (ABC) and Instrumental Forest (ABC) refer to a causal forest/instrumental forest built using ABC and estimating effects using EHS covariates. For forest methods, we calculate standard errors using bootstrapping.

As we did for ITT, we can also account for different population density of heterogeneity-relevant observables by using [Athey et al. \(2019\)](#)’s instrumental causal forest approach to estimate the LATE of EHS, had EHS been implemented in a population like ABC’s.⁶ This allows us to further the comparability of the estimates by estimating what would be the LATE that we would estimate for EHS if the joint distribution of heterogeneity-relevant covariates in EHS was the same as it was in ABC. Using the estimated (instrumental) causal forest, we can easily estimate what that LATE of EHS would look like under ABC’s density of relevant covariates. The results are presented in the LATE panel, Row 6, Column 3. We estimate the instrumental forest using the same heterogeneity-relevant covariates discussed in the ITT analysis to build the trees in the forest. The results are similar, with a LATE of 1.47 instead of 1.44 for any participation, and 1.33 instead of 1.10 for the 12-month participation definition.

3.5 Attrition and Lee Bounds

As discussed above, attrition is a nontrivial concern. For the center-only programs, IQ is missing for around 36% of children (Table [A.1](#)). Attrition is also unbalanced by treatment status: the age-three IQ response rates are 73% for treated children and 55% for control children. However, as we show next, our main conclusion—that the effects of ABC and EHS are comparable once we align populations and program types—is robust to this missing-data problem.

We take three complementary steps to assess the potential consequences of this differential attrition. First, Tables [A.3–A.5](#) compare baseline characteristics between treated and control children conditional on being observed at age three. We conduct these tests for our full vector of covariates and separately across the three program types. Reassuringly, none of the differences are statistically significant, and the estimated differences do not systematically favor either the treatment or control group. While balance on observables cannot rule out differential selection on unobservables, the absence of systematic differences across a rich set of baseline characteristics makes substantial differential selection less plausible.

Second, we assess whether restricting the sample to children with observed IQ scores changes estimated treatment effects on the Achenbach behavioral-problems scale, an outcome collected through the Parent Interview and available for a substantially larger sample. The estimated effects do not differ statistically significantly when we restrict the sample to children with observed IQ scores (Appendix Table [A.6](#)), providing additional evidence against substantial differential selection into the IQ sample.

⁶See also [Angrist and Fernandez-Val \(2010\)](#) on LATE extrapolation to other populations without using machine learning techniques.

Third, we estimate Lee bounds for all of our main results. This approach is particularly relevant because response rates are higher in the treatment group. Lee bounds require monotonicity in the effect of treatment assignment on the probability of being observed, an assumption that is plausible in this setting. The original EHS evaluation identifies two channels through which treated families were more likely to remain in follow-up: greater accessibility to the survey team through their continued connection to EHS programs, particularly at center-based sites, and greater willingness to continue participating in the evaluation. In particular, the evaluation documents that some control-group families refused further participation after learning that they would not receive EHS services (Love et al., 2001, p. B.11).

Table 4 presents the results across a range of specifications. In our preferred specification, which estimates the LATE for center-only programs in the subsample comparable to ABC, the bounds range from 0.49 to 3.26. This interval lies strictly above the upper bounds for the ITT estimates in the full sample and in the broader subsample, and its lower bound remains economically large. The interval also encompasses the tighter ABC bounds of 0.88 to 1.37. Taken together, these results indicate that, even under conservative attrition assumptions, restricting attention to comparable populations and center-based services yields sizable EHS impacts of an order of magnitude similar to those of the ABC program.

Table 4: Lee Bounds for EHS and ABC

ITT						
Program	EHS					ABC
Type	All	Center + Mixed	Center Only			
Full	(-0.04, 0.28)	(0.03, 0.35)	(-0.18, 0.52)			(0.89, 1.07)
Subsample	(0.03, 0.34)	(0.13, 0.39)	(0.23, 0.68)			(0.83, 1.17)
LATE						
Program	EHS					ABC
Type	All	Center + Mixed	Center Only			
Participation	Any	Any	Any	6m	12m	12m
Full	(-0.17, 1.46)	(0.08, 1.03)	(-0.26, 0.98)	(-0.26, 1.00)	(-0.26, 1.06)	(0.95, 1.26)
Subsample	(0.14, 2.17)	(0.35, 1.72)	(0.49, 3.26)	(0.44, 2.73)	(0.42, 2.55)	(0.88, 1.37)

Notes: Participation: “Any” refers to participating in any part of the program, whereas 6m and 12m refer to participating in at least 6 or 12 months of the program, respectively. The subsample excludes non-black children and highly educated mothers.

4 Complier Subgroups and subLATEs

4.1 subLATEs

The LATE parameter discussed in the last section allows us to learn something about the effect of actual EHS participation, not just the *offer* of participation. While this certainly provides more information relative to ITT, it still commingles two quite distinct parameters that could be of ultimate interest: the effect of participating in EHS relative to keeping the child at home, and the effect of participating in EHS relative to having the child attend whatever alternative child care arrangement the child might be sent to, when not participating in EHS. This second effect is conceptually related to earlier literature (e.g. [Heckman et al. \(2000\)](#)) that considers close program substitutes available to subjects randomized out of a program.

In our case, other ECE programs were available to both ABC and EHS participants⁷. So, the untreated group in the EHS evaluation includes children who stay at home and children who attend other centers. Therefore, in our context, individuals can comply in more than one way when not assigned to an EHS offer. In all cases, these compliers do not participate in EHS when they do not receive the offer, but some keep the child at home, while others send the child to an alternative child care center. In what follows, we denote the former as “nh”-compliers and the latter as “ch”-compliers.

A growing literature has recognized the importance of distinguishing between these various complier subgroups. In principle, the two treatment effects discussed above can be defined separately for these two complier subgroups. [Kline and Walters \(2016\)](#) and [Feller et al. \(2014\)](#) carefully analyzed this issue in the case of Head Start, which serves 3- to 5-year-olds. Following [Kline and Walters \(2016\)](#), we can then represent the “overall” LATE of EHS as a weighted average of two subLATEs, one associated with “nh”-compliers and one associated with “ch”-compliers, where the relative prevalences of “nh” and “ch” compliers act as weights:

$$\text{LATE} = \omega_{nh}\text{nh-LATE} + \omega_{ch}\text{ch-LATE} \quad (1)$$

where the subLATEs are given by:

$$\text{nh-LATE} = E[Y_{i,h} - Y_{i,n} | i \text{ is nh-complier}]$$

⁷See [García et al. \(2020\)](#) on the impact of accounting for program substitutes when analyzing the long-run effects of ABC.

$$\text{ch-LATE} = E[Y_{i,h} - Y_{i,c} | i \text{ is ch-complier}].$$

nh-LATE is the average treatment effect of attending the assigned program (h) instead of keeping the child at home (n), for the nh-complier subpopulation that has the child attending the program when randomly given an offer to attend, but keeps the child at home when randomly denied such an offer. The ch-LATE is defined analogously as the average effect of participating in EHS relative to attending an alternative center, for the subpopulation of “ch”-compliers who switch into EHS when receiving an offer but send their child to an alternative center when an EHS offer is not received. The weights are given by $\omega_{nh} = \frac{p_{nh}}{p_{nh} + p_{ch}}$ and $\omega_{ch} = \frac{p_{ch}}{p_{nh} + p_{ch}}$, where p_{nh} and p_{ch} are the relative prevalences of nh- and ch-compliers in the EHS population. The “overall” LATE is then a weighted average of two subLATEs whose magnitudes we cannot separately identify: the nh-LATE and the ch-LATE. However, we identify and describe the weights in the weighted average below in Table 5.

Identifying these subLATEs is not straightforward. For example, as [Kline and Walters \(2016\)](#) did, one could use the structure of a Roy model to identify them. However, as [Fang et al. \(2023\)](#) argue, without this additional structure, these subLATEs are typically only partially identified. Moreover, a Roy model approach, as in [Kline and Walters \(2016\)](#), would require an additional instrument beyond the randomized EHS offer. We, therefore, consider bounds for these subLATEs following [Berkes et al. \(2024\)](#). Even restricting ourselves to only bounds requires us to make the assumptions that $\text{ch-LATE} \leq \text{nh-LATE}$ and $\text{ch-LATE} \geq 0$. The first assumption is natural if we think the impact of EHS participation relative to staying at home is at least as large as the impact of attending alternative preschools relative to staying at home. This need not be the case, so we just take this assumption only as a benchmark to compute these bounds. The assumption is plausible, though. While alternative child care centers may be higher quality than EHS, the centers the “Subsample” population can access are likely no higher than EHS.

From Equation (1), because the “overall” LATE is a weighted average of the two subLATEs, that “overall” LATE represents the upper bound for ch-LATE and the lower bound for nh-LATE. According to the assumption, the lower bound for ch-LATE would be 0.⁸ Given the uncertainty regarding the assumption underlying this last bound, we also present values of ch-LATE that are slightly below 0.

⁸This assumption also implies an upper bound for nh-LATE. Assuming that $\text{ch-LATE} = 0$ is the same as assuming that the impact of the alternative centers and the impacts of EHS are the same. In that case, we can find the impacts by redefining the participation variable as attending any center (alternative or EHS). The impact of randomization on this newly defined variable will now include only the nh-compliers, as the ch-compliers will be considered participating regardless of their randomization status. Then, we can obtain the LATE for that new participation variable as ITT/p_{nh} . That will be the upper bound for nh-LATE.

The estimated bounds are reported in Table 5 and are shown visually as the light-orange shaded area in Figure 1a. For comparison, we also report the bounds for the corresponding subLATEs in ABC as the green shaded area. While the bounds on the size of ch-LATE are similar for EHS and ABC, the nh-LATE upper bound for EHS is much higher than for ABC. The figure shows that, because the LATEs for both programs are similar, the range of subLATEs for ABC falls within the possible range of subLATEs for EHS. Next, we investigate the relationship between these estimates.

Table 5: subLATE Bounds and Compliance Shares for EHS and ABC

Program	EHS Center Only	ABC		
	Prevalence of Compliance Types			
Observations	110	98		
Compliers				
p_{nh}	0.19	0.61		
p_{ch}	0.24	0.27		
Share				
ω_{nh}	0.44	0.69		
Always-Takers				
p_{hh}	0.11	0.06		
p_{cc}	0.26	0.02		
p_{nn}	0.20	0.04		
subLATE Bounds				
	ch-LATE	nh-LATE	ch-LATE	nh-LATE
Bounds				
Lower Bound	0.00	1.11	0.00	1.06
Upper Bound	1.11	2.52	1.06	1.54

Notes: Sample is restricted to black children with low maternal education. EHS participation is defined as 12 months.

Rearranging Equation (1), we can express one subLATE as a linear function of the other subLATE, the “overall” LATE, and the relative complier subgroups’ shares $\omega = (\omega_{nh}, \omega_{ch})$:

$$\text{ch-LATE} = \frac{\text{LATE}}{\omega_{ch}} - \left(\frac{\omega_{nh}}{\omega_{ch}} \right) \text{nh-LATE}. \quad (2)$$

Note that both the “overall” LATE and the complier subgroup shares in Equation (2) are identified. To identify the complier subgroups, we need only information on the choice regarding the child’s participation (EHS, Alternative Center, or None) and EHS offer assignment. With those, we can estimate the prevalence of the various complier and always-taker types from the following system of identifying equations:

$$\begin{aligned}
\Pr(D = h|Z = 1) &= p_{hh} + p_{nh} + p_{ch} \\
\Pr(D = c|Z = 1) &= p_{cc} \\
\Pr(D = n|Z = 1) &= p_{nn} \\
\Pr(D = h|Z = 0) &= p_{hh} \\
\Pr(D = c|Z = 0) &= p_{cc} + p_{ch} \\
\Pr(D = n|Z = 0) &= p_{nn} + p_{nh}
\end{aligned} \tag{3}$$

where, in addition to p_{ch} and p_{nh} which we introduced above, we now also consider the prevalence of various always-taker groups: those who always participate in EHS (p_{hh}), those who always participate in an alternative center (p_{cc}), and those who are really never-takers in the sense that they always stay at home and never participate in EHS nor in an alternative center (p_{nn}). Table 5 presents the estimated prevalences for each complier type that solve (3). As expected, the nh- and ch-complier shares are much larger in ABC (61% and 27%) than in EHS (19% and 24%) because the non-compliance problem is much smaller in ABC. Correspondingly, the shares of various always-taker types (hh, cc, nn) are much larger in EHS (11%, 26%, 20%) than in ABC (6%, 2%, 4%). More importantly, the *relative* difference in compiler shares is quite different in EHS versus ABC. In EHS, only 44% of the compilers are nh-compilers, whereas in ABC, 69% of all the compilers are of the “nh” type.

With the estimates of ω in hand, we can plot the relationship between the possible values of the two EHS subLATEs as given in Equation (2). This linear relationship is reported as the solid orange line in Figure 1a. If the assumptions required to compute the bounds hold, we know the relationship between the two subLATEs will hold somewhere within the shaded area. The solid line within the shaded area shows the possible pairs of the two subLATEs consistent with the two point-identified objects: the “overall” LATE and the relative complier shares (ω_{nh}, ω_{ch}). Naturally, the solid lines are downward sloping because there is an inverse relationship between the two in the sense that when one of the subLATEs is larger, the other one must be smaller to remain consistent with the “overall” LATE.

Figure 1a shows that, given any level of the nh-LATE, the values of the ch-LATE for EHS are higher than those of ABC. For the lower values of the nh-LATE, which are more

plausible, the values of the ch-LATE are more similar across both programs.

4.2 Accounting for Differences in Relative Complier Subgroup Shares

An alternative way to compare the two programs is to construct a counterfactual LATE for EHS by taking the weighted average of the two EHS subLATEs, but using the “complier weights” from ABC’s population, ω^{ABC} , instead. This accounts for differences in relative complier subgroup shares across the two programs. The complier shares are partly endogenous to program quality, but they are also strongly influenced by the supply of alternative programs. By using ABC’s complier weights, we adjust for the greater abundance of substitutes during EHS. Indeed, as shown in Table 5, the nh-complier share in EHS is 0.44, while ABC’s is 0.69. The distinction between absolute and relative prevalence across complier subgroups is important here. While the share of ch-compliers was relatively similar between EHS and ABC (24% vs 27%), the share of nh-compliers is much larger in the ABC population, thus explaining a much larger relative complier share of home complier ω_{nh} in the ABC population. Suppose that for both programs the nh-LATE is much larger than the ch-LATE but the sizes of the two subLATEs were the same across both programs (so the programs are equally effective). One would then still find that EHS underperforms relative to ABC just because ABC’s LATE puts much more weight on the larger subLATE than EHS. We can use the bound for the EHS subLATEs weighted by the ABC subLATE weights to avoid this problem. This gives the LATE of EHS on cognition if the EHS population had the same relative prevalence of complier subtypes as ABC. To that end, one can define LATE_{ABC}^{EHS} , the local average treatment effect of EHS participation on cognition using relative compliance shares from ABC ($\omega_{nh}^{ABC}, \omega_{ch}^{ABC}$):

$$\text{LATE}_{ABC}^{EHS} = \omega_{nh}^{ABC} \text{nh-LATE}^{EHS} + \omega_{ch}^{ABC} \text{ch-LATE}^{EHS}. \quad (4)$$

Inserting Equation (2) into Equation (4) yields:

$$\text{LATE}_{ABC}^{EHS} = \omega_{nh}^{ABC} \text{nh-LATE}^{EHS} + \omega_{ch}^{ABC} \frac{\text{LATE}^{EHS} - \omega_{nh}^{EHS} \text{nh-LATE}^{EHS}}{\omega_{ch}^{EHS}} \quad (5)$$

$$= \frac{\omega_{ch}^{ABC}}{\omega_{ch}^{EHS}} \text{LATE}^{EHS} + \left(\omega_{nh}^{ABC} - \frac{\omega_{ch}^{ABC} \omega_{nh}^{EHS}}{\omega_{ch}^{EHS}} \right) \text{nh-LATE}^{EHS}. \quad (6)$$

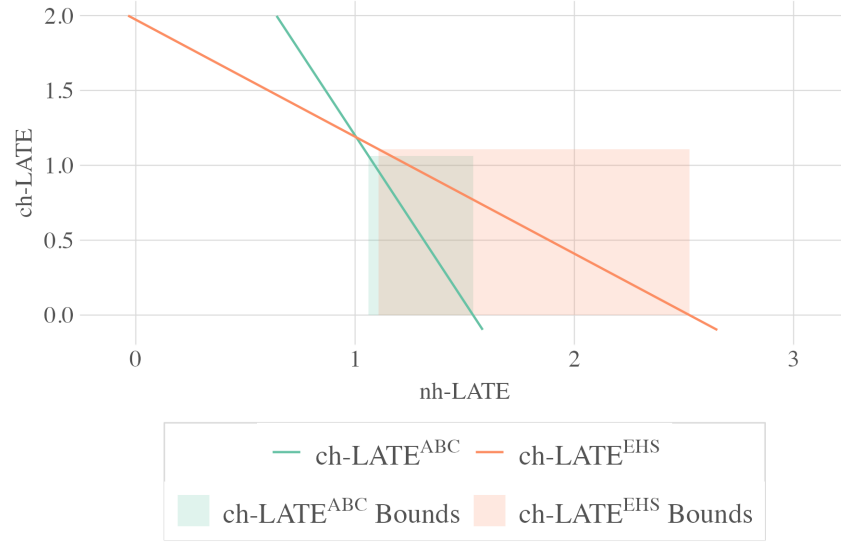
Therefore, the counterfactual “what if” LATE_{ABC}^{EHS} can be expressed as a linear function

of the *unknown* nh-LATE^{EHS} but with a *known* intercept $\frac{\omega_{ch}^{ABC}}{\omega_{ch}^{EHS}} \text{LATE}^{EHS}$ and *known* slope,

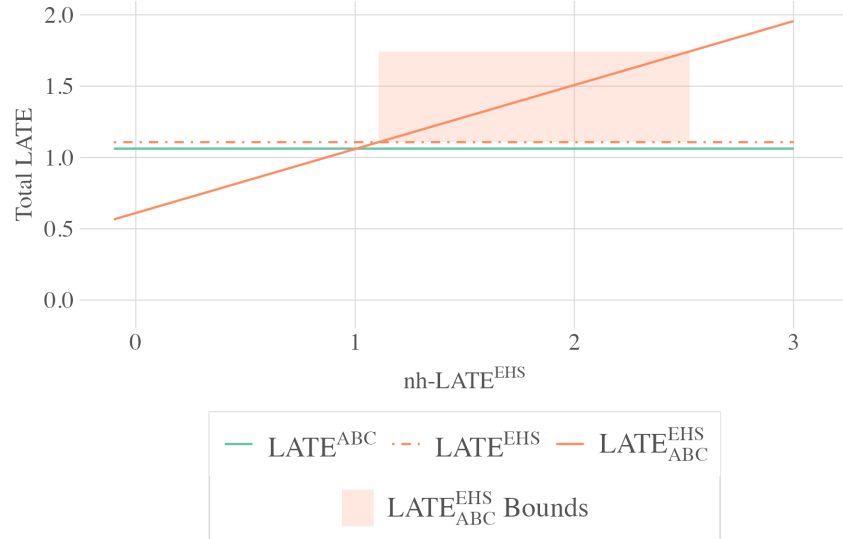
$$\frac{\omega_{nh}^{ABC} \omega_{ch}^{EHS} - \omega_{ch}^{ABC} \omega_{nh}^{EHS}}{\omega_{ch}^{EHS}},$$

where the sign of the slope depends on the sign of $\omega_{nh}^{ABC} \omega_{ch}^{EHS} - \omega_{ch}^{ABC} \omega_{nh}^{EHS} = \omega_{nh}^{ABC}(1 - \omega_{nh}^{EHS}) - (1 - \omega_{nh}^{ABC})\omega_{nh}^{EHS} = \omega_{nh}^{ABC} - \omega_{nh}^{EHS}$. The slope is positive because, as discussed above, $\omega_{nh}^{ABC} > \omega_{nh}^{EHS}$. We graph this linear function over a range of nh-LATE values. To approximate the most relevant range, we take as a reference the bounds for this subLATE derived in Subsection 4.1. The results are depicted in Figure 1b. In the likely range for EHS's nh-LATE given by the bounds, the counterfactual total LATE for EHS would always be higher than the estimated 1.11 (dashed orange line), and the larger the nh-LATE is, the more so. This is because the total LATE calculation now puts more weight on the nh-LATE. This analysis implies that, if the composition of counterfactual alternatives had been the same for both programs, the overall LATE for EHS would have exceeded that of ABC. The magnitude of this gap would increase as the difference between the nh-LATE and the ch-LATE grows.

Figure 1: ch-LATE^{EHS} and LATE_{ABC}^{EHS} as functions of nh-LATE^{EHS}
(EHS Center-Based Programs Only - Subsample)



(a) Bounds for nh-LATE and ch-LATE



(b) Bounds for LATE_{ABC}^{EHS}

Notes: Sample is restricted to black children with low maternal education and with program participation defined using 12 months. Top Panel: Green and orange shades represent the bounds for nh-LATE and ch-LATE, where ch-LATE is bounded below at 0 and above at LATE, and nh-LATE is bounded below at LATE. Bottom Panel: The solid green line represents the LATE for ABC, and the dashed orange line represents the LATE for EHS. The solid orange line represents the linear relationship for LATE using the weights from ABC. The orange shade represents the bounds for nh-LATE and the resulting LATE_{ABC}^{EHS} , where nh-LATE is bounded below at LATE.

5 External Validity

We have shown that, when comparisons are made appropriately, the impacts estimated in the randomized evaluation of EHS are similar in magnitude to those of ABC. A central question, however, is whether these impacts are externally valid and can inform policy. Following [List \(2020\)](#), we consider how our evidence meets key conditions for generalizability.

The natural population of interest for our study is children 0-3 who participate in Early Head Start nationally. The National Early Head Start Research and Evaluation (EHSRE) Study enrolled participants between 1996 and 1998 in 17 programs drawn from the first two waves of 143 federally funded programs. These programs were selected to provide a broad national distribution: they covered every region of the country, both urban and rural areas, and included all major EHS service approaches. Families served at the sites reflected the demographic and socioeconomic diversity typical of Early Head Start nationally. While strong research partnerships influenced site selection, they did not participate in program implementation. Importantly, program operations were not altered for research purposes: aside from over-recruitment to allow for random assignment, sites followed their usual enrollment procedures, and no exemptions were granted to excuse families from participation ([Love et al., 2005](#)). We have already discussed attrition in detail in Section 3.5. Compliers account for 43% of the sample (Table 5), a large and policy-relevant share of families most likely to be affected by program expansion. Our final sample, which selects the subgroup most comparable to ABC, focuses on center-based programs, which account for about 65% of EHS services,⁹ and on Black families without a college degree, a group of particular policy relevance.

A key strength for external validity is that the EHSRE took place in a natural program environment, where both recruitment and service delivery functioned as fully scaled-up public programs, not as a pilot or demonstration. Families made participation decisions and experienced services under the same norms, constraints, and incentives that characterize standard program operation. A caveat is that the study was conducted in the early years of EHS, so implementation quality varied. Still, at the time of the evaluation, EHS had already been implemented at scale, in more than 140 sites nationally. ([Love et al., 2005](#)) report that a few of the sites ranked low on an implementation rating scale. This suggests current EHS impacts may be even higher than those we report, as programs have benefited from decades of institutional learning and professionalization.

Taken together, these features show that the EHSRE mirrors the diversity of EHS programs nationwide, ensuring our findings rest on a representative foundation and inform

⁹<https://headstart.gov/program-data/article/head-start-program-facts-fiscal-year-2023>

national policy.

6 Conclusions

We revisit Abecedarian (ABC), a legendary early childhood education program, and the federal scaled-up program it inspired, Early Head Start (EHS). A randomized evaluation of EHS was commissioned at the program’s launch in the mid-1990s. Early assessments of EHS effectiveness from that evaluation were, at best, lukewarm, as the program appeared not to live up to the sizeable impacts witnessed years earlier for its predecessor, ABC. These somewhat disappointing results may have contributed to EHS’s limited expansion in the three decades since its inception.

This apparent “voltage drop” as the early childhood education program scaled up has been singled out as a poster child for the idea, now well established in academic and policy circles, that many programs that show great promise in the pilot or demonstration phase fail to deliver the same impacts when implemented at scale. A growing literature now examines why so many programs do not scale up effectively (see e.g., [List \(2021\)](#) and the references cited therein).

Our work complements this literature arguing that once careful comparisons between EHS and ABC are made, EHS appears to be as effective as, or even *more* effective than, ABC. Our findings have important implications for the cost-benefit analysis of potential EHS expansions. In the paper, we discuss in detail which adjustments drive most of the narrowing in the gap in reported estimates. In sum, we show that focusing on EHS programs most similar to ABC and adjusting for differences in the populations served almost eliminates the effectiveness gap. Moreover, accounting for differential non-compliance suggests that participation in EHS had stronger treatment effects than participation in ABC. Moreover, based on our partial identification results for subLATES, when we consider the effectiveness of EHS had it faced the ABC population, we find that the advantage of EHS over ABC becomes stronger.

We also argue that our LATEs commingle subLATEs that might be of ultimate interest. We provide bounds for these subLATEs that isolate the effects of EHS relative to a particular fallback option when an EHS offer is not received: either stay at home or attend an alternative center. Drawing on our analysis of the subLATEs, we go on to argue that our estimates finding strong EHS effects are conservative in that the ultimate comparison, which involves evaluating the effectiveness of participating in EHS relative to staying at home, is likely more highly diluted for EHS than for ABC. This is because by the mid-1990s, when the EHS was evaluated, the child care industry had already matured in the United States,

providing many more high-quality alternatives for children not attending Early Head Start.

References

- Achenbach, T. M. and L. A. Rescorla (2016, August). Achenbach System of Empirically Based Assessment. Institution: American Psychological Association.
- Al-Ubaydli, O., M. S. Lee, J. A. List, and D. Suskind (2021, September). The Science of Using Science A New Framework for Understanding the Threats to Scaling Evidence-Based Policies. In *The Scale-Up Effect in Early Childhood and Public Policy*, Volume 7.
- Angrist, J. and I. Fernandez-Val (2010, December). ExtrapoLATE-Ing: External Validity and Overidentification in the LATE Framework. Working Paper 16566, National Bureau of Economic Research.
- Armor, D. J. (2014, October). The Evidence on Universal Preschool: Are Benefits Worth the Cost? SSRN Scholarly Paper 2549381, Social Science Research Network, Rochester, NY.
- Athey, S., J. Tibshirani, and S. Wager (2019, April). Generalized Random Forests. *The Annals of Statistics* 47(2), 1148–1178.
- Berkes, J., A. Bougen, D. Filmer, and T. Fukao (2024). Improving Preschool Provision and Encouraging Demand: Evidence from a Large-Scale Construction Program. *Journal of Public Economics*.
- Besharov, D. J., P. Germanis, and C. A. Higney (2011, September). Early Head Start Research and Evaluation Project. In *Assessments of Twenty-Six Early Childhood Evaluations*, Volume 7, pp. 21. Maryland School of Public Policy.
- Bloom, H. S. (1984). Accounting for No-Shows in Experimental Evaluation Designs. *Evaluation Review* 8(2), 225–242.
- Currie, J. and D. Almond (2011, January). Human Capital Development Before Age Five. In D. Card and O. Ashenfelter (Eds.), *Handbook of Labor Economics*, Volume 4, pp. 1315–1486. Elsevier.
- Dunn, L. M. (1965). *Expanded Manual Peabody Picture Vocabulary Test*. Minneapolis, MN: American Guidance Service, inc.
- Elango, S., J. L. García, J. J. Heckman, and A. Hojman (2015, November). Early Childhood Education. Working Paper 21766, National Bureau of Economic Research.

- Fang, Z., A. Santos, A. M. Shaikh, and A. Torgovitsky (2023). Inference for Large-Scale Linear Systems with Known Coefficients. *Econometrica*.
- Feller, A., T. Grindal, L. Miratrix, and L. Page (2014). Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care-Type Setting. Technical report, Harvard University.
- García, J. L., J. J. Heckman, D. E. Leaf, and M. J. Prados (2020, July). Quantifying the Life-Cycle Benefits of an Influential Early-Childhood Program. *Journal of Political Economy* 128(7), 2502–2541.
- Heckman, J., N. Hohmann, J. Smith, and M. Khoo (2000). Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment. *The Quarterly Journal of Economics* 115(2), 651–694.
- Imbens, G. W. and J. D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica: Journal of the Econometric Society*, 467–475.
- Kline, P. and C. R. Walters (2016, November). Evaluating Public Programs with Close Substitutes: The Case of Head Start. *The Quarterly Journal of Economics* 131(4), 1795–1848.
- Lim Brodowski, M. and S. F. Naom (2021). Enabling Contexts to Support Scale-Up Lessons from Early Head Start–Child Care Partnerships. In *The Scale-Up Effect in Early Childhood and Public Policy*.
- List, J. (2021). *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*.
- List, J., D. Suskind, and L. H. Supplee (2021). *The Scale-Up Effect in Early Childhood and Public Policy. Why Interventions Lose Impact at Scale and What We Can Do About It*. Routledge.
- List, J. A. (2020). Non Est Disputandum De Generalizability? A Glimpse into the External Validity Trial. Technical report, National Bureau of Economic Research.
- List, J. A. (2026). *Experimental Economics: Theory and Practice*. University of Chicago Press.
- List, J. A., I. Muir, and G. Sun (2024). Using Machine Learning for Efficient Flexible Regression Adjustment in Economic Experiments. *Econometric Reviews* 44(1), 2–40.

- Love, J. M., E. E. Kisker, C. Ross, H. Raikes, J. Constantine, K. Boller, J. Brooks-Gunn, R. Chazan-Cohen, L. B. Tarullo, C. Brady-Smith, A. S. Fuligni, P. Z. Schochet, D. Paulsell, and C. Vogel (2005). The Effectiveness of Early Head Start for 3-Year-Old Children and Their Parents: Lessons for Policy and Programs. *Developmental Psychology* 41(6), 885–901.
- Love, J. M., E. E. Kisker, C. M. Ross, P. Z. Schochet, J. Brooks-Gunn, K. Boller, D. Paulsell, A. S. Fuligni, and L. J. Berlin (2001, June). Building Their Futures: How Early Head Start Programs Are Enhancing the Lives of Infants and Toddlers in Low-Income Families: Volume II: Technical Report Appendixes. Technical report, Mathematica Policy Research, Princeton, NJ.
- Love, J. M., E. E. Kisker, C. M. Ross, P. Z. Schochet, J. BrooksGunn, D. Paulsell, K. Boller, J. Constantine, C. Vogel, A. S. Fuligni, and others (2002). Making a Difference in the Lives of Infants and Toddlers and Their Families: The Impacts of Early Head Start. Volumes III: Final Technical Report.
- Moffitt, R. (2006). Forecasting the Effects of Scaling Up Social Programs: An Economics Perspective.
- Ramey, C. T. (1974). The Carolina Abecedarian Project: A Longitudinal and Multidisciplinary Approach to the Prevention of Developmental Retardation. Technical report.
- Ramey, C. T., D. M. Bryant, J. J. Sparling, and B. H. Wasik (1985, July). Project CARE: A Comparison of Two Early Intervention Strategies to Prevent Retarded Development. *Topics in Early Childhood Special Education* 5(2), 12–25.
- Sparling, J. J. (1974). Synthesizing Educational Objectives for Infant Curricula. *Annual Meeting of the American Educational Research Association*.
- Terman, L. M. and M. A. Merrill (1960). *Stanford-Binet Intelligence Scale*. Cambridge MA: The Riverside Press.
- United States Department Of Health And Human Services. Administration For Children And Families (2004). Early Head Start Research and Evaluation (EHSRE) Study, 1996-2010: [United States].
- Wasik, B. H., C. T. Ramey, D. M. Bryant, and J. J. Sparling (1990, December). A Longitudinal Study of Two Early Intervention Strategies: Project CARE. *Child Development* 61(6), 1682–1696.

Supplemental Appendix

Appendix: Cognitive Skill Measurements

In the two programs analyzed in this paper, we measure cognitive ability. We standardize them to make the scales more comparable. In EHS, we use the Peabody Picture Vocabulary Test-Third Edition (PPVT). PPVT is a receptive vocabulary test that measures auditory comprehension of spoken words in children and adults. This test is administered individually to people ages 2 to 90+ and lasts 10 to 15 minutes. The PPVT raw score is the total number of correct responses, including the unasked questions below the basal item and above the ceiling item. The raw score is thus equal to the total number of errors subtracted from the number of the ceiling question (Dunn (1965), p. 10). In ABC, we use the Stanford-Binet Intelligence Scale (SB). The SB revises the Binet-Simon Scale, the first scale to use the concept of mental age and intelligence quotient as the ratio of mental age to chronological age (Terman and Merrill (1960), p. 5–6). The test is a collection of tasks whose items, in aggregate, measure general intelligence.

Appendix: Covariates

The covariates included in some of the analyses are: mother’s education (less than high school, completed high school or GED, and college and above), mother’s IQ, mother’s age, child’s race (whether black or not), number of siblings, gestational age, whether the child has a father figure at home, and child’s sex.

Appendix: Sample Selection

The EHS analysis sample is selected as follows. The initial EHS sample, which includes all program types, included $N = 2,977$ observations. After excluding programs with visits only (i.e., keeping center-only and mixed program types), the sample is reduced to $N = 1,609$. We then exclude observations with missing outcome, participation, or covariate data, further reducing the sample to $N = 768$. For most of the analysis, we focus on “center only” programs, with a sample of $N = 375$. We also work with a subsample of $N = 110$ EHS study participants in “center only” sites that exclude non-black children and highly educated mothers (the same “subsampling” also reduces the ABC sample from 103 to 98). Table A.1 provides additional detail on the sample selection process.

Table A.1: Analysis Sample Selection

Program	EHS						ABC	
Type	All		Center + Mixed		Center Only			
Sample	Full	Subsample	Full	Subsample	Full	Subsample	Full	Subsample
All	2,977 (100%)	757 (100%)	1,609 (100%)	482 (100%)	610 (100%)	165 (100%)	113 (100%)	107 (100%)
Non-Missing Outcome	1,424 (48%)	376 (50%)	831 (52%)	257 (53%)	392 (64%)	115 (70%)	104 (92%)	99 (93%)
Non-Missing Covariates	1,382 (46%)	376 (50%)	817 (51%)	257 (53%)	386 (63%)	115 (70%)	104 (92%)	99 (93%)
Non-Missing Participation (Any)	1,374 (46%)	375 (50%)	811 (50%)	257 (53%)	384 (63%)	115 (70%)	103 (91%)	98 (92%)
Non-Missing Participation (Center)	1,267 (43%)	348 (46%)	768 (48%)	245 (51%)	375 (61%)	110 (67%)	103 (91%)	98 (92%)
Non-Missing Alternate Care	1,267 (43%)	348 (46%)	768 (48%)	245 (51%)	375 (61%)	110 (67%)	103 (91%)	98 (92%)

Notes: EHS Full sample includes center only, home visit only, and mixed. ABC only had center-based care. Non-Missing Participation (Any) excludes those missing any participation information. Non-Missing Participation (Center) excludes those missing center participation information. The subsample excludes non-black children and highly educated mothers. Each row inherits and expands upon the previous restriction.

Appendix: Additional Results

Table A.2: Balance Conditional on Complete Estimation Data: ABC

Sample	Full			Subsample		
	T	C	$p(\Delta)$	T	C	$p(\Delta)$
Children's Characteristics						
% Male	49.1%	46.0%	0.76	46.0%	47.9%	0.85
% Black	96.2%	100.0%	1.00	100.0%	100.0%	1.00
# Siblings	0.5	0.8	0.14	0.5	0.8	0.24
Gestational Age (Weeks)	39.2	39.8	0.22	39.1	39.8	0.24
Mother's Characteristics						
Mother's IQ	84.4	83.5	0.71	83.4	83.3	0.98
Mother's Age	19.3	20.2	0.37	19.3	20.0	0.44
% HS Completed	35.8%	22.0%	0.12	38.0%	22.9%	0.11
% College Completed	3.8%	4.0%	0.95	0.0%	0.0%	1.00
% Father Figure at Home	20.8%	34.0%	0.14	20.0%	33.3%	0.14
% Above Poverty	15.5%	8.0%	0.23	12.5%	8.3%	0.50
Sample Size						
# Observations	53	50		50	48	

Notes: Mothers' years of education are measured at age 0 for ABC and at age 2 for EHS. Mother's education is categorized as less than high school, completed high school or GED, and college and above. Child's race has multiple categories, which we collapse to a binary on being black. ABC reports direct income measures, and we construct the poverty binary using the federal poverty line.

Table A.3: Balance Conditional on Complete Estimation Data: EHS - All

Sample	Full			Subsample		
	T	C	$p(\Delta)$	T	C	$p(\Delta)$
Children's Characteristics						
% Male	49.2%	48.8%	0.90	47.3%	50.6%	0.53
% Black	38.7%	36.0%	0.31	100.0%	100.0%	1.00
# Siblings	1.0	1.0	0.94	1.0	1.1	0.71
Gestational Age (Weeks)	39.3	39.4	0.44	39.5	39.4	0.43
Mother's Characteristics						
Mother's IQ	90.6	90.6	0.95	85.2	85.2	0.94
Mother's Age	22.1	22.0	0.88	20.1	20.5	0.52
% HS Completed	31.6%	30.9%	0.78	39.0%	38.0%	0.84
% College Completed	28.6%	26.8%	0.48	0.0%	0.0%	1.00
% Father Figure at Home	20.3%	23.0%	0.23	6.6%	5.4%	0.65
% Above Poverty	12.1%	10.7%	0.46	9.3%	5.4%	0.16
Sample Size						
# Observations	671	596		182	166	

Notes: Mothers' years of education are measured at age 0 for ABC and at age 2 for EHS. Mother's education is categorized as less than high school, completed high school or GED, and college and above. Child's race has multiple categories, which we collapse to a binary on being black. ABC reports direct income measures, and we construct the poverty binary using the federal poverty line.

Table A.4: Balance Conditional on Complete Estimation Data: EHS - Center + Mixed

Sample	Full			Subsample		
	T	C	$p(\Delta)$	T	C	$p(\Delta)$
Children's Characteristics						
% Male	50.1%	48.8%	0.71	46.9%	52.1%	0.41
% Black	44.1%	42.5%	0.64	100.0%	100.0%	1.00
# Siblings	1.0	0.9	0.18	1.2	1.2	0.91
Gestational Age (Weeks)	39.3	39.3	0.86	39.5	39.4	0.75
Mother's Characteristics						
Mother's IQ	89.2	89.4	0.77	83.7	84.4	0.47
Mother's Age	22.2	21.9	0.45	19.9	20.3	0.56
% HS Completed	28.9%	30.9%	0.55	34.4%	33.3%	0.86
% College Completed	29.9%	25.9%	0.22	0.0%	0.0%	1.00
% Father Figure at Home	18.0%	19.6%	0.57	7.0%	5.1%	0.53
% Above Poverty	13.3%	10.7%	0.27	10.9%	6.8%	0.26
Sample Size						
# Observations	405	363		128	117	

Notes: Mothers' years of education are measured at age 0 for ABC and at age 2 for EHS. Mother's education is categorized as less than high school, completed high school or GED, and college and above. Child's race has multiple categories, which we collapse to a binary on being black. ABC reports direct income measures, and we construct the poverty binary using the federal poverty line.

Table A.5: Balance Conditional on Complete Estimation Data: EHS - Center Only

Sample	Full			Subsample		
	T	C	$p(\Delta)$	T	C	$p(\Delta)$
Children's Characteristics						
% Male	50.2%	47.4%	0.60	50.0%	47.7%	0.82
% Black	42.4%	38.8%	0.49	100.0%	100.0%	1.00
# Siblings	0.8	0.8	0.42	0.7	0.5	0.38
Gestational Age (Weeks)	39.5	39.4	0.50	39.8	39.5	0.37
Mother's Characteristics						
Mother's IQ	88.7	89.0	0.80	82.7	84.9	0.13
Mother's Age	22.5	22.4	0.88	20.9	22.5	0.21
% HS Completed	30.1%	32.1%	0.69	40.9%	43.2%	0.82
% College Completed	31.1%	25.6%	0.25	0.0%	0.0%	1.00
% Father Figure at Home	16.9%	21.2%	0.31	6.1%	9.1%	0.57
% Above Poverty	12.8%	10.9%	0.58	10.6%	9.1%	0.79
Sample Size						
# Observations	219	156		66	44	

Notes: Mothers' years of education are measured at age 0 for ABC and at age 2 for EHS. Mother's education is categorized as less than high school, completed high school or GED, and college and above. Child's race has multiple categories, which we collapse to a binary on being black. ABC reports direct income measures, and we construct the poverty binary using the federal poverty line.

Appendix: Impacts on Child Behavior

Table A.6: Treatment Effects on Achenbach Scores

Program Type	EHS		
	All	Center + Mixed	Center Only
	ITT		
Full/No Conditioning	-0.06 (0.04)	-0.07 (0.06)	-0.09 (0.09)
Full/Has IQ	-0.09* (0.05)	-0.14* (0.07)	-0.16 (0.10)
Full/ p -value: Has IQ = No Conditioning	0.289	0.081	0.215
Subsample/No Conditioning	-0.20** (0.09)	-0.21* (0.11)	-0.20 (0.17)
Subsample/Has IQ	-0.16 (0.10)	-0.22* (0.12)	-0.25 (0.18)
Subsample/ p -value: Has IQ = No Conditioning	0.366	0.931	0.401
	LATE		
Full/No Conditioning	-0.09 (0.06)	-0.11 (0.08)	-0.16 (0.15)
Full/Has IQ	-0.12* (0.07)	-0.19* (0.10)	-0.26 (0.17)
Full/ p -value: Has IQ = No Conditioning	0.356	0.104	0.228
Subsample/No Conditioning	-0.31** (0.13)	-0.38** (0.19)	-0.52 (0.45)
Subsample/Has IQ	-0.24 (0.15)	-0.38* (0.22)	-0.64 (0.46)
Subsample/ p -value: Has IQ = No Conditioning	0.339	0.966	0.483
Sample Size: Full/No Conditioning	2,031	1,101	451
Sample Size: Full/Has IQ	1,367	789	378
Sample Size: Subsample/No Conditioning	487	332	129
Sample Size: Subsample/Has IQ	368	255	113

Notes: Achenbach System of Empirically Based Assessment (ASEBA) assesses competencies, strengths, adaptive functioning, and behavioral, emotional, and social problems from age 1.5 to 90+ years ([Achenbach and Rescorla \(2016\)](#)). The scores are standardized, and larger scores indicate behavioral problems. “No Conditioning” only restricts to non-missing Achenbach scores. “Has IQ” restricts to having both non-missing Achenbach and IQ scores. “ p -value” reports p -value for the test that the “No Conditioning” and “Has IQ” estimates are the same.