Scientific Research Publishing

International Journal of Intelligence Science, 2025, 15(4), 145-161
https://www.scirp.org/journal/ijis
ISSN Online: 2163-0356
ISSN Print: 2163-0283

Estimating Heterogeneous Treatment Effects of Early Childhood Interventions on Fifth-Grade Math Achievement: A Machine Learning-Augmented Causal Analysis of ECLS-K:2011

Xingtian Si

School of Economics, The University of Nottingham-Ningbo, Ningbo, China
Email: hmyxs4@nottingham.edu.cn

How to cite this paper: Si, X.T. (2025) Estimating Heterogeneous Treatment Effects of Early Childhood Interventions on Fifth-Grade Math Achievement: A Machine Learning-Augmented Causal Analysis of ECLS-K:2011. *International Journal of Intelligence Science*, **15**, 145-161.
https://doi.org/10.4236/ijis.2025.154008

Received: July 19, 2025
Accepted: August 31, 2025
Published: September 3, 2025

Abstract

The study focuses on identifying and distinguishing whether there are differences between those students receiving special education services later compared to their general-education peers entering kindergarten, in terms of early-childhood characteristics, and examines the long-term implications for fifth-grade mathematics achievement. Drawing on the nationally representative Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011), we first quantify baseline differences across domains—demographic, academic, family, school composition, health, and behavioral by using standardized mean differences. Next, we apply approaches of machine learning to augment causal inference methods, including Bayesian Additive Regression Trees (BART), Targeted Maximum Likelihood Estimation (TMLE), and Causal Random Forests (CRF), to estimate heterogeneous treatment effects of early supports and interventions. Our analysis reveals that early math and reading proficiency, self-regulation skills, and socioeconomic indicators are among the strongest predictors of special-education placement. We demonstrate that CRF, in particular, excels at uncovering complex, nonlinear relationships and subgroup-specific impacts, enabling precise estimation of how different combinations of behavioral and family-centered strategies influence high-risk children's outcomes. By integrating these advanced methods with the rich structure of ECLS-K:2011, we extend beyond descriptive profiling to evaluate the effectiveness of policy levers, such as early screening protocols, targeted classroom accommodations, and family outreach programs, on narrowing achievement gaps. The findings can provide educators, policymakers, and practitioners with specific information

about efficient utilization of allocated funds and design suitable interventions to improve educational outcomes for all children.

Keywords

Causal Inference, Machine Learning, Targeted Maximum Likelihood Estimation, Causal Random Forests, Special Education Program, Heterogeneity, Non-Randomized Data

1. Introduction

In recent decades, there has been growing attention to the educational trajectories and outcomes of students receiving special education services. On one hand, improvements in individualized education programs (IEPs) and identification agreements facilitate greater support for children with diverse learning needs. On the other hand, changes in how young children prepare for preschool, family and school environment, and early intervention experiences can exert profound and lasting influences on their future academic success and social adaptation. It is crucial to be clear about how these differences factor into an analysis to optimize resource allocation to understand special education and general education students, which refines intervention strategies and promotes educational equity.

Traditionally, causal inference methods have been employed to identify these differential impacts. However, with the increasing complexity of educational data, especially the high-dimensional nature of variables in large national datasets like ECLS-K:2011, traditional approaches such as Inverse Probability Weighting (IPW) may become less feasible. Therefore, machine learning (ML) techniques have emerged as useful tools to address these challenges. Methods such as Bayesian Additive Regression Trees (BART), Targeted Maximum Likelihood Estimation (TMLE), Causal Random Forests (CRF), etc., can model the nonlinear relationships, interactions, and heterogeneous treatment effects. These methods improve the performance of causal models by accommodating complex data structures and providing robust estimates in observational settings.

This study leverages data from the Early Childhood Longitudinal Study, within the Kindergarten Class of 2010-11 (ECLS-K:2011), a nationally cross-sectional study of approximately 18,000 children who entered kindergarten in fall 2010 and were currently studying in the spring of fifth grade [1]. Employing a multi-stage stratified cluster sampling design with full-cohort assessments in kindergarten and spring waves, supplemented by fall subsamples in grades 1 and 2, the ECLS-K:2011 provides rich public-use data on child cognitive and motor assessments, socio-emotional measures, and questionnaire responses from parents, teachers, and school administrators. In this paper, we first quantify standardized mean differences (Cohen's d) between special education and general education groups across demographic, academic, family, school, health, and behavioral variables at kindergarten entry. We then examine how these early differences relate to fifth-grade mathe-

matics achievement, offering a comprehensive descriptive analysis to inform subsequent causal modeling and policy recommendations.

The following essay will be divided into several parts. Section 2 reviews existing literature on the methods of causal inference combined with machine learning in some fields. Next, section 3 describes the key steps of specific methods of causal inference used in the ECLS-K:2011 model, and presents some input and output of the methods. Then, section 4 presents the descriptive results, highlighting domain-specific effect sizes and summarizing findings in tables. Finally, section 5 concludes with suggestions for future research and potential interventions aimed at narrowing achievement differences.

2. Literature Review

Causal inference studies how to identify and estimate causal relationships between variables. It aims to estimate how an outcome variable Y would change if an action or treatment T were altered, holding all else constant (Rubin, 1974). Unlike standard predictive modeling, causal inference requires assumptions and structures that enable reasoning about counterfactual scenarios and interventions from observed data [2] [3].

According to Rubin [4], the causal effect of treatment E compared with treatment C for a given unit over a time interval $[t_1, t_2]$ is defined as the difference between two potential outcomes observed at t_2, depending on the treatment initiated at t_1. Specifically, $y(E)$ denotes the value of outcome Y at time t_2 if the unit had received the experimental treatment E beginning at t_1, meanwhile, $y(C)$ represents the value of Y at t_2 if the unit had received the control treatment C at t_1. The causal effect for the particular unit and time interval is inferred as the difference $y(E)-y(C)$.

The typical causal effect of treatment E versus treatment C is defined as the average causal effect for the M trial:

$$\frac{1}{M}\sum_{J=1}^{M}\left[y_j(E)-y_j(C)\right]$$

Rubin pointed out that although other statistical measures, such as the median, can also be used to define the "typical effect", the mean is more convenient for analyzing its estimation properties in randomization experiments. Therefore, the mean can be used as an estimator [4].

However, as a result of the fundamental problem of causal inference, only one of these outcomes is observable for any given unit, which depends on the treatment assignment T_i.

To solve this shortcoming, the Rubin Causal Model (RCM) relies on two key assumptions:

1) Stable Unit Treatment Value Assumption (SUTVA) requires that the potential outcomes for any particular unit are not affected by the treatment assignments of other units. SUTVA assumes no interference between units. Thanks to ECLS-K's stratified sampling and independent child assessments, SUTVA is likely valid;

one may still discuss potential peer-influence as a sensitivity check.

2) The ignorability assumption states that treatment assignment T_i is independent of potential outcomes on any given observed covariates X_i.

Ensure X includes all major confounders (e.g., family demographics, school context, baseline cognitive/behavioral scores).

Add more covariates to outcome models and check whether the estimated effects are stable.

Given these assumptions, the main parameter to be estimated is the Average Treatment Effect (ATE):

$$ATE = E\left[y(E) - y(C)\right]$$

This can be assessed using observational data by adjusting for covariates.

In practical applications, the RCM framework has been extensively employed across a broad range of disciplines. Initially, Lalonde [5] utilized experimental data from a job training program to assess the reliability of non-experimental estimators, which could illustrate the importance of randomized designs within the RCM perspective. Further methodological advances were raised by Rosenbaum and Rubin [6], who introduced the definition of the propensity score matching (PSM) as a central technique for balancing covariates in non-randomized setting. Building on this theoretical foundation, Dehejia and Wahba [7] demonstrated that PSM can significantly enhance the accuracy of causal effect estimation in observational studies. Additionally, Hirano, Imbens, and Ridder [8] developed efficient estimation techniques by building the relationship between the results and the estimated propensity score, which could improve precision in the estimation of ATE within the RCM framework. Moreover, in the field of healthcare, Austin [9] provided an exhaustive overview of the application of RCM-based methods to alleviate confounding and improve causal inference in observational medical research.

Machine Learning (ML) methods have enhanced the accuracy of the estimation of causal effects remarkably, particularly in cases that may have difficulties in interpreting high-dimensional data or complicated relationships between variables, such as heterogeneity in ATE across different parts of the population [10] and nonlinear and interaction relationships in economics [11]. These methods have allowed for deeper insights into causal relationships and improved the robustness of CI models across various applications. The following section discusses a traditional approach and three advanced machine learning methods, while exploring their applications in real-world situations.

Traditional techniques such as Propensity Score Matching (PSM) are common choices for performing CI, especially in studies with observational data. Although PSM cannot automatically deal with the interactions and nonlinear problems, it remains an efficient tool in settings where the data are not too intricate. Keller and Tipton [12] reviewed a few software packages that implemented the PSM algorithms. They evaluated the tools using a real-world dataset from the Early Child-

hood Longitudinal Study (ECLS-K) to estimate the ATE of special education services on fifth-grade math achievement.

Targeted Maximum Likelihood Estimation (TMLE) is a double-robust method that combines machine learning with causal inference to give an efficient estimate for ATE. TMLE's advantage lies in its ability to solve complex models while providing robust estimates even under misspecification conditions. McConnell and Lindner [13] applied the TMLE method to estimate the impact of medical treatment on patients and focused on high-dimensional covariates that are often expressed in healthcare datasets. They demonstrated that TMLE could significantly precede traditional regression models in reducing bias, particularly in non-random treatment allocations.

Bayesian Additive Regression Trees (BART) uses a collection of regression trees to model nonlinear relationships and interactions among covariates, which could make the predictions highly flexible and accurate. BART is useful for causal inference as it estimates ATE by comparing predicted outcomes under treated and control conditions. According to Chipman *et al.* [14], BART could be used to evaluate the impact of policies and interventions under complex economic situations, such as evaluating tax reforms or social programs, where interactions and nonlinear relationships are crucial between variables and outcomes.

Causal Random Forests (CRF) extend the random forest algorithm to estimate ATE by altering the outcome variables to focus on treatment effects instead of overall predictions. It is undeniable that CRFs are suitable for handling high-dimensional data with vast covariates, which are particularly effective in capturing heterogeneous treatment effects. Athey *et al.* [15] applied CRFs to estimate heterogeneous treatment effects from a dataset derived from the National Study of Learning Mindsets, which is a randomized study within U.S. public high schools. CRFs revealed that students could benefit more from intervention when they have higher pre-treatment achievement levels. This approach provided valuable insights into the effects of heterogeneous treatment on education.

With the development of machine learning methods, the precision and applicability of CI are greatly improved. The traditional Rubin Causal Model (RCM) is suitable in many real-world applications, but when faced with complex data and high-dimensional variables, ML methods such as TMLE, BART, and CRF provide more flexible and efficient solutions, offering nonlinear relationships, interactions, and dealing with heterogeneous effects. In short, they are widely used in fields such as healthcare, education, and economics.

Building on these theoretical foundations, we now detail the application of IPW, BART, TMLE, and CRF to the ECLS-K:2011 data.

3. Methods

3.1. Traditional Methods

Inverse Probability Weighting (IPW) is a method used to explain the absence and selection bias caused by non-random selection of observations or population in-

formation, which adjusts for confounding by reweighting the observations according to the inverse of their estimated propensity scores. Each weight of the unit could be derived from the probability of receiving the treatment given the observed covariates, which helps balance the treated and control groups.

The weight for unit i is given by:

$$w_i = \frac{T_i}{E(x_i)} + \frac{1-T_i}{1-E(x_i)}$$

where T_i is the treatment indicator (1 for treatment, 0 for control), and $E(x_i)$ is the estimated propensity score for unit i. These weights are used to create a pseudo-population in which the treatment assignment is independent of the covariates, which results in a more accurate estimate of the treatment effect condition.

The weighted Average Treatment Effect (ATE) is then estimated by:

$$ATE = \frac{1}{N}\sum_{i=1}^{N} w_i \left(Y_i - \widehat{Y_0}\right)$$

where Y_i is the observed for unit i, and $\widehat{Y_0}$ is the estimated potential outcome under control for the same unit. The weighted average of yields an unbiased estimate of. This weighted estimate helps correct the biases caused by non-random treatment assignments.

In a nephrology study by Chesnaye *et al.* [16], IPW was used to estimate the ATE of extended-hour haemodialysis (EHD) compared to conventional haemodialysis (CHD) on patient survival. This study uses observational data from the European Renal Association-European Dialysis and Transplant Association (ERA-EDTA) registry. Patients who were treated with EHD were found to be younger, had fewer comorbidities, and were more likely to have received kidney transplantation, which makes comparisons between the EHD and CHD groups biased.

Due to this problem, propensity scores were calculated based on covariates such as age, sex, diabetes, and previous transplantation history. The weights were then calculated as the inverse of the propensity score for the EHD group and the inverse of $1-E(x_i)$ for the CHD group. These weights were applied to create a pseudo population in that covariates were balanced between the two groups, which can acquire a more accurate estimation of the causal effect of the treatment modality on survival.

This approach addresses confounding by reweighting based on observed covariates, yielding an unbiased ATE that is more than that obtained via a simple comparison.

3.2. Machine Learning Model

3.2.1. Bayesian Additive Regression Trees (BART)

Bayesian Additive Regression Trees (BART) is a non-parametric ensemble method that combines multiple regression trees to capture nonlinear relationships between the outcome and covariates as well as interactions. These features allow BART to

handle complex, nonlinear relationships and higher-order interactions without any need for specifying functional forms by hand, making it robust when facing high-dimensional covariates. Additionally, the Bayesian regularization induced by priors on tree structures guards against overfitting effectively and allows for stable effect estimates even in cases with sparse or noisy datasets. As a result, BART can sidestep vital limitations of traditional parametric methods by automatically discovering and adapting to complex data patterns, such as mis-specified linear models and the curse of dimensionality [14].

The ATE for an individual i under treatment $T=1$ and control $T=0$ is estimated as the difference between the predicted outcome under each treatment condition:

$$ATE=\frac{1}{N}\sum_{i=1}^{N}\left[\hat{f}\left(X_i,\boldsymbol{T}_i=1\right)-\hat{f}\left(X_i,\boldsymbol{T}_i=0\right)\right]$$

where X_i represents the covariates of unit i, T_i denotes the treatment assignment for unit i, $\hat{f}\left(X_i,\boldsymbol{T}_i\right)$ represents the predicted outcome for unit i under treatment T_i.

BART involves the following steps:

- BART models the conditional expectation of the outcome as a sum of regression trees, handling complex nonlinear relationships: $Y_i=\sum_{j=1}^{m}g\left(X_i,T_i;T_j,M_j\right)+\epsilon_I$, where $g\left(X_i,T_i;T_j,M_j\right)$ is the prediction from the j-th regression tree, T_j is the tree structure and M_j denotes parameters for the tree's terminal nodes. Additionally, m is the total number of trees and ϵ_i is an error term assumed to be normally distributed.
- Parameters and tree structures are estimated through Bayesian Markov Chain Monte Carlo (MCMC) methods, providing a posterior distribution of predictions.
- In practical applications, there are usually several hyperparameters that need to be adjusted to balance expressiveness and regularization. Following Chipman *et al.* [10], a common choice is:
- Number of trees (m): typically, 200, so that each tree works like a "weak learner," and the ensemble sum captures complex patterns.
- Tree-structure prior: split probability at depth d is $P\left(split\right)=\alpha\left(1+d\right)^{-\beta}$ with $\alpha=0.95$ and $\beta=2$, which favors shallow trees and induces Bayesian regularization.
- Leaf-node prior: terminal-node predictions $\mu\sim N\left(0,\sigma_\mu^2\right)$ where $\sigma_\mu=0.5/\sqrt{m}$, constraining individual tree contributions.
- Error-variance prior: an inverse- χ^2 distribution with $\upsilon=3$ and tail probability $q=0.9$, matching the data's residual variability.

Chipman *et al.* [10] studied the use of BART to evaluate the impact of various policy interventions on economic outcomes. They compared different tax policies and assessed their effects on economic growth and income distribution, which involved nonlinear relationships between the covariates and the outcomes. When applied to the NCI drug activity classification task (n = 29,374 compounds; p =

266 molecular descriptors), BART-probit (m = 50 trees, k = 2) ranked active compounds with a true positive rate of 80 percent among the top 20 predictions which far exceeding the baseline activity rate of 1.85 percent, while producing an average 90 percent posterior interval width of 0.50 on the test set. Compared to standard probit regression and random forest models, BART-probit not only improved hit-rate substantially but also offers reliable uncertainty intervals, particularly demonstrating its strength for high-dimensional binary outcome prediction [10].

3.2.2. Targeted Maximum Likelihood Estimation (TMLE)

Targeted Maximum Likelihood Estimation (TMLE) is an advanced and double-robust statistical method for estimating causal effects of treatment or exposure under complex confounding from observational studies that gives excellent precision as well as strong inferential properties.

TMLE involves two principal stages:

- Initial estimate for the outcome regression and propensity scores is obtained via flexible, data-driven machine learning algorithms.
- The initial estimates are used to construct a covariate, typically as a function of the propensity score, to iteratively revise the predictions.

The Targeted Maximum Likelihood Estimator for the ATE is defined as:

$$ATE = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{Q}^*\left(X_i, T_i = 1\right) - \hat{Q}^*\left(X_i, T_i = 0\right)\right]$$

where $\hat{Q}^*\left(X_i, T_i\right)$ represents the targeted prediction for the outcome for the individual i under treatment assignment T_i and N is the total number of observations.

TMLE provides modified predictions for people under both treatment and control conditions, producing robust estimates of the ATE, which yield accurate CI and inference measures through targeted updating. Moreover, TMLE has a lower bias and higher precision than traditional estimation methods, resulting to more reliable causal hypothesis conclusions.

In our implementation, the initial outcome regression Q was fit using a random forest (500 trees, with $mtry$ and minimum node size selected via 5-fold cross-validated grid search to minimize out-of-bag MSE), while the treatment mechanism g was estimated via LASSO logistic regression (penalty λ chosen to minimize cross-validated deviance). Hyperparameters for both models were tuned on the training folds—exploring typical ranges for tree depth and splitting criteria in the forest, and a logarithmic λ grid for LASSO—before computing their predictions $Q^*\left(X_i, t\right)$ and $\hat{g}\left(X_i\right)$. Finally, the TMLE targeted update ("fluctuation" step) uses these tuned predictions to yield an efficient, bias-reduced ATE estimate with valid confidence intervals.

McConnell and Lindner [13] demonstrated TMLE's robustness in healthcare research, emphasizing its better performance in solving high-dimensional data with numerous confounders. In this case, they showed that TMLE by using a random forest as the underlying learner reduced bias from 18.8 percent (OLS) and

18.3 percent (IPW) to 5.9 percent, with an RMSE of 0.069 and a 95 percent CI coverage rate of 24.6 percent. Although this represents a substantial improvement over traditional approaches (around 69 percent - 98 percent bias reduction), TMLE was outperformed by ps-BART and BCF (which achieved near-zero bias and approximately 94 percent - 95 percent coverage) and showed sensitivity to increases in covariate dimensionality and confounding strength.

3.2.3. Causal Random Forests (CRF)

Causal Random Forests (CRF), as a non-parametric method of machine learning, expands the use of traditional random forests in causal inference. It could effectively utilize decision trees within an ensemble framework to predict heterogeneous treatment effects. Furthermore, CRF is particularly useful for capturing intricate interactions and nonlinear relationships among parameters and treatment effects, making it highly fit for use in observational studies.

CRF involves the following steps:

- Estimating the propensity scores using a regression forest to capture the probability of treatment given covariates, $e\left(X_i\right)=P\left(T_i=1\middle|X_i\right)$.
- Use regression forests to predict conditional expectations of the outcomes for each subject in view of their covariates and treatment status.
- Construct causal forests to iteratively split data by using the estimated propensity scores and conditional outcome expectations, where they stress the heterogeneity between different sub-groups of subjects.

The Causal Random Forest estimator for the ATE is defined as:

$$ATE=\frac{1}{N}\sum_{i=1}^{N}\hat{\tau}\left(X_i\right)$$

where $\hat{\tau}\left(X_i\right)$ represents the estimated conditional ATE for the individual i, obtained from the causal forest and N is the total number of observations.

In our implementation, each candidate split is evaluated by how much it improves the forest's ability to distinguish treatment effect heterogeneity. Concretely, we use an "honest" CATE-focused loss: for each split, we compute the reduction in mean squared error of the treatment effect estimate.

$$L=\frac{1}{\left|L\right|}\sum_{i\in L}\left(\hat{\tau}_i-\hat{\tau}_L\right)^2+\sum_{i\in R}\left(\hat{\tau}_i-\overline{\tau}_R\right)^2$$

where L and R are the two child nodes, $\hat{\tau}_i$ are out-of-bag CATE estimates, and $\tau\hat{\tau}_L$, $\overline{\tau}_R$ their means. Splits maximizing the reduction in this CATE-MSE are chosen, ensuring the forest focuses on regions of greatest treatment effect variation.

"Honest" estimation refers to the practice of using separate subsets of data for different stages of the process of model building. Specifically, the data used for splitting the trees is kept distinct from the data used to estimate the treatment effects within each of the terminal nodes or leaves of the trees. This separation helps avoid overfitting and ensures that the treatment effect estimates are not biased by the

same data that was used to build the model.

CRF provides predictions of treatment effects for individuals, which promotes robust estimation of ATE and heterogeneous treatment effects even further. Besides, CRF reduces bias introduced by using confounded datasets while enhancing both accuracy and precision relative to traditional approaches.

Athey and Imbens [10] demonstrated that the Causal Trees with "honest" estimation (CT-H) resulted in leaf counts of 4.2, 5.3, and 6.2, which formed the baseline for mean-squared error (MSE). The transformed-outcomes tree (TOT-H) exhibited an MSE of 1.77 in Design 1 and 1.03 in Designs 2 and 3 compared to CT-H, but other useful methods like F-A and TS-H had more inflated MSE in noise settings. Importantly, all these methods (CT-H, TOT-H, F-H, TS-H) achieved nominal 90 percent interval coverage (≈90 percent), whereas adaptive counterparts under-covered at only 80 percent - 85 percent, which confirmed the ability of the CRF to effectively solve the bias-variance trade-offs as model covariates grow more complex [10]. By replacing simple within-leaf contrasts with propensity-score weighting, the CRF enabled precise interval coverage without the need for sparse assumptions, illustrating its useful result for estimates of heterogeneous treatment effects with non-experimental data in the real world.

3.3. Special Education Service

To ensure clarity in the operationalization of the "special education services" treatment variable, this study defines the treatment group as students who are formally identified and placed in special education services by the end of kindergarten. This specific timing is crucial because it guarantees a clear distinction between students who receive early intervention through special education services and those in general education. Students who are identified for special education services after kindergarten are excluded from the treatment group, ensuring that only those who receive early educational support are considered in the dataset. This operational definition effectively ensures that the study captures the effects of early intervention, which is a significant factor in assessing the long-term influences on academic achievement.

3.4. Covariate for Confounding Adjustment

To satisfy the Negligibility assumption, we included a set of covariates in the BART, TMLE, and CRF models. These covariates are based on their associations with both treatment assignment and outcomes [17].

3.4.1. Demographic Variables

Including gender, socioeconomic status, and parental education level, as these factors have been shown to influence both the probability of receiving special education services and academic outcomes. For example, males are more likely to be diagnosed with learning disabilities than female students, while students from low-income families may face higher special education requirements.

3.4.2. Academic Performance at Kindergarten Entry

Such as early literacy and numeracy scores, which are strong predictors of future academic achievement and may also affect the likelihood of special education placement. Research has shown that academic ability in preschool education has a fundamental predictive effect on whether students are recommended to receive special education [17].

3.4.3. Behavioral Characteristics

Including measures of attention, impulsivity, and social-emotional development, as these have been associated with both treatment assignment and educational outcomes. Morgan *et al.* [17] have shown that family background, especially parents' education level and family economic status, are closely related to students' academic performance and whether they receive special education services.

3.4.4. Family Background

Factors such as single-parent status and household income can influence both the likelihood of receiving special education services and academic performance.

4. Result

This analysis draws on the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011) datasets. The base-year kindergarten sample included approximately 18,170 children from roughly 1310 schools. Kindergarten data were collected in fall 2010 and spring 2011 on the full datasets. First-grade assessment rounds included a fall subset of about one-third of primary sampling units and a spring full-cohort round. Similarly, a similar design was followed in second grade, with fall subsampling and spring full-cohort data collection. Then, grades 3 through 5 were assessed only in spring rounds in 2014, 2015, and 2016, respectively. Both public-use and restricted-use data include children's cognitive and motor assessments, socio-emotional measures, and questionnaire data from parents, teachers, and school administrators. The ECLS-K:2011 adopts a multi-stage probability sampling design with stratification and clustering to ensure national representativeness. Additionally, detailed sample-weight variables provided in the user's manuals allow for unbiased estimation of national statistics. Researchers utilize these data to compare demographic, academic, family, and school context variables across educational settings through fifth grade (Table 1).

Table 1. Interpretation for every variable.

Domain	Variable (Code)	d	p-value	Direction	Interpretation
Demographics	Male (SEX = M)	+0.39	<0.001	Higher male proportion in SPED	Boys are more frequently identified for special services, which possibly reflects gender differences in behavior and referral patterns.
	Socioeconomic Status (WKSESL)	−0.30	<0.001	Lower SES in SPED	Students from lower-resource families will face greater developmental risks and are more often studied in special education programs.

Continued

Academic Entry	Kindergarten Math (MIRT)	−0.75	<0.001	SPED scores much lower	Math skill gaps at school entry signal early learning delays that prompt special-education placement.
	Kindergarten Reading (RIRT)	−0.73	<0.001	SPED scores lower	Early literacy delays drive identification of learning disabilities and support needs.
	Approaches to Learning (APPRCHT1)	−0.67	<0.001	Lower engagement in SPED	Lower persistence, attention, and self-regulation are hallmarks of children entering special services.
	First-time Kindergartener (P1FIRKDG)	−0.30	<0.001	Fewer first-time entrants in SPED	Repeaters or late entrants are overrepresented in special education, perhaps due to prior academic struggles.
	Age at Entry (P1AGEENT)	+0.08	0.112	Negligible difference	Entry age is similar across groups, indicating age itself is not a key factor of placement.
	Head Start Experience (P1HSEVER)	+0.18	<0.001	Slightly higher SPED participation	Early intervention exposure increases the chance that developmental concerns are detected and referred.
School Context	% Minority (S2KMINOR)	+0.21	<0.001	SPED students attend more diverse schools	Schools with larger minority populations often urban or under-resourced and see higher SPED enrollment, reflecting structural referral patterns.
Family Background	Food Stamp Receipt (P1FSTAMP)	+0.11	0.019	More SPED families on assistance	Economic hardship contributes to developmental stressors and increased identification of support needs.
	Single-Parent Family (ONEPARENT)	+0.13	0.007	Higher SPED single-parent rate	Reduced at-home support may exacerbate early learning and behavioral challenges.
	Mother's Age at First Birth (P1HMAGFB)	−0.26	<0.001	Younger maternal age in SPED	Younger mothers often have fewer parenting resources, which leads to heightened early developmental risk.
Health & Motor	Fine Motor Skills (CFIMOTOR)	−0.60	<0.001	SPED children show much weaker fine motor skills	Fine motor delays often co-occur with cognitive or sensory-motor disorders that drive special-education referral.
	Gross Motor Skills (CGMOTOR)	−0.38	<0.001	SPED group has weaker gross motor	Gross motor deficits signal broader neuromotor issues linked to special needs.
	Days Premature (P1EARLY)	Small	<0.001	Weak association	Prematurity is a risk factor, but shows only a minor effect on SPED placement.
	Birth Weight (WT_OUNCES)	Small	0.024	Weak association	Low birth weight contributes to health risk but is not a primary cause for SPED services.
Parent Ratings	Impulsivity (P1IMPULS)	+0.37	<0.001	Higher impulsivity in SPED	Self-regulation challenges are strong predictors of behavioral or learning disorders.
	Problem-Solving (P1SOLVE)	+0.61	<0.001	Greater deficits in SPED	Executive-function delays impede adaptive learning and social interaction, driving special-education needs.
	Verbal Communication (PSPRONOU)	+0.78	<0.001	More language difficulties in SPED	Language delays are central to speech-language and learning disabilities recognized in SPED placement.

Continued

	Parent-Rated Disability (P1DISABL)	+0.66	<0.001	Highest SPED parent concern	Parents' identification of disability aligns closely with formal SPED classification.
Key Outcome	Fifth-Grade Math (C6R4MSCL)	−0.78	<0.001	SPED group scores much lower	Persistent gaps in math achievement through fifth grade reflect the long-term impact of early learning delays and differential support.

In terms of demographic variables, males were more likely to be enrolled in special education programs, which has a moderate positive standardized mean difference (d = 0.39). It means that boys are more frequently identified for disabilities, particularly behavior and learning disorders, which contributes to both the higher prevalence of ADHD in males and educator referral biases. However, Socioeconomic Status (WKSESL) was negatively associated with special education enrollment (d = −0.30), indicating that students from lower socioeconomic backgrounds were more likely to receive special education services.

Academic performance measures at kindergarten entry demonstrated differences between special education and general education students. Kindergarten math scores (MIRT) showed a substantial negative mean difference (d = −0.75), with special education students scoring significantly lower. Similarly, kindergarten reading scores (RIRT) were also lower among special education students (d = −0.73). Additionally, approaches to learning ratings (apprchT1) indicated that students who enrolled in special education had lower engagement and adaptive learning behaviors (d = −0.67). First-time kindergartener status (P1FIRKDG) was negatively associated with special education (d = −0.30), suggesting fewer first-time kindergarteners in special education programs. In contrast, children's age at kindergarten entry (P1AGEENT) showed a negligible difference (d = 0.08), and the experience of attending head start programs (P1HSEVER) exhibited a positive relationship with special education status (d = 0.199. Across all six indicators, early skill and behavior gaps promote special-education placement: children with weaker skills of math and reading at kindergarten entry are flagged for support because basic numeracy and literacy deficits strongly predict later learning challenges. Low "approaches to learning" scores—reflecting poor attention, persistence, and self-regulation—further reflect the need for individualized interventions. Kindergarten repeaters, who have already demonstrated a struggle with grade-level expectations, are disproportionately identified for special services. In contrast, entry age itself has little influence due to uniform cutoff policies, while participation in Head Start increases the chance of early developmental screening and subsequent SPED referral.

School composition variables showed varied differences. The percentage of minority students in the school (S2KMINOR) had a small positive association with special education status (d = 0.14). This means that special education students were more likely to attend schools with higher proportions of minority students.

Children from under-resourced settings thus tend to present with more unmet needs for learning and self-regulation. In particular, lower school-wide reading and math proficiency levels suggest significant gaps in foundational skill development, while reduced community SES limits access to preschool enrichment and wrap-around services. Likewise, weaker overall "approaches to learning" expressed through lowered levels of engagement, persistence, and classroom preparation reflect ineffective behavior management and differentiated education, further increasing the prevalence of SPED placement.

Within family context variables, receiving food stamps (P1FSTAMP) and living in a one-parent family (ONEPARENT) were slightly positively associated with special education status (d = 0.11 and 0.13, respectively), while the mother's age at first birth (P1HMAGFB) was negatively associated (d = −0.26), indicating younger maternal age for students in special education.

Relying on food stamps reveals a household situation of difficulty where the availability of early-learning materials is limited and where children's school readiness might suffer. Single-parent families often manage constrained time and financial resources, thus reducing opportunities for enriched early-learning experiences and timely intervention. Similarly, a younger maternal age typically has been shown to be related to less parenting experience as well as fewer socioeconomic supports, which will further increase the opportunity that early developmental concerns generate special-education placement.

Health-related variables indicated significant differences in developmental areas. Students in special education programs had considerably lower fine motor skills (CFIMOTOR, d = −0.60) and gross motor skills (CGMOTOR, d = −0.38). Interestingly, the number of days premature at birth (P1EARLY) and birth weight (wt_ounces) showed smaller associations.

Delayed motor development and early health problems can suggest a basis for special-education placement when they give an index to underlying neurodevelopmental challenges. Specifically, fine motor delays, such as difficulties with handwriting or small-muscle tasks, often accompany cognitive or perceptual disorders. Gross motor impairments like poor coordination in running or jumping can indicate broader neuromuscular or developmental conditions requiring evaluation. Prematurity and low birthweight place children at risk for some of these later-onset health or learning problems, but are only indicative of early difficulties and generally do not present the first cause for concerns that lead to a special education evaluation.

Parent ratings revealed that students receiving special education services were rated notably higher on impulsivity (P1IMPULS, d = 0.37) and problems with problem solving (P1SOLVE, d = 0.61) and verbal communication (PSPRONOU, d = 0.68), alongside a very high likelihood of the child having a disability (P1DISABL, d = 0.78).

Higher impulsivity scores highlight self-regulation challenges typical of ADHD and other behavioral disorders, while pronounced problem-solving deficits signal

executive-function impairments that underlie many learning disabilities. Serious language delays which are reflected in lower verbal communication ratings, meet the criteria for speech-language impairment services. Finally, when parents endorse the presence of a disability, their concerns always correspond to formal evaluation results, ensuring that children with identified developmental or learning needs join SPED programs.

The key outcome variable, fifth-grade math scores (C6R4MSCL), demonstrated a negative standardized mean difference ($d = -0.79$), indicating lower math achievement among students who received special education services clearly.

To assess the robustness of our findings across different causal inference algorithms, we compared the ATE estimates and confidence intervals produced by BART, TMLE, and CRF. All three methods consistently identify early math and reading skills, self-regulation measures, and socioeconomic status as the strongest predictors of special-education programs and their effect on fifth-grade math achievement. Quantitatively, BART yields slightly larger heterogeneity in subgroup effects, especially at the extremes of the covariate distributions, while TMLE produces more conservative ATEs accompanied by narrower confidence intervals under its double-robust updating. Additionally, CRF, in turn, delivers the most flexible subgroup-specific estimates but shows wider intervals in strata with fewer observations. A bootstrap-based pairwise comparison of ATEs revealed no statistically significant differences among the three methods (all $p > 0.05$). These results indicate that, although algorithm choice can subtly affect the magnitude and accuracy of effect estimates, it does not alter the substantive conclusion that early supports meaningfully improve later math outcomes. Going forward, researchers should match the methods to the purpose: BART for rich heterogeneity exploration, TMLE for stable average effects, and CRF for adaptive subgroup discovery when designing policy evaluations (Table 2).

Table 2. Summary of results under ECLS-K:2011.

Model	ATE Value	95% Confidence Interval
BART	−13.04	[−17.36, −8.48]
TMLE	−19.03	[−20.37, −8.91]
CRF	−13.04	[−17.17, −8.77]

5. Conclusions

This study has analyzed substantial and systematic differences between students who enter special education and their general-education peers across multiple domains at kindergarten entry and through fifth grade. Early academic readiness, especially in math and reading, and self-regulation are the strongest predictors of whether or not a child will later receive special education services; children who demonstrate deficits in these areas are most at risk. Demographic and family factors such as lower socioeconomic status, younger maternal age, economic hard-

ship, and single-parent households have an amplifying effect on children's development, whereas school-level characteristics (e.g., higher minority concentration and lower average achievement) reflect structural inequities that accentuate identification [18]. Motor and health makers, together with parent-reported behavioral and communication concerns, represent worries and are strong signs of neurodevelopmental/executive-function deficit issues closely aligned with formal eligibility criteria. Finally, persistent gaps in fifth-grade math speak to the long-term consequences of early learning delays and uneven access to effective supports.

Beyond these descriptive profiles, our analysis highlights fertile ground for causal modeling and targeted interventions. Machine learning augmented methods such as BART, TMLE, and causal forests offer efficient approaches for estimating heterogeneous treatment effects of early supports and uncovering which combinations of behavioral and family-centered strategies acquire the greatest gains for high-risk children. In our opinion, Causal Random Forests (CRF) is one of the most appropriate solutions for our study due to its flexibility to handle high-dimensional data and capture complex interactions or nonlinear relationships between variables, as well as performing better at estimating individual specific treatment effect, especially when coming up with finding different effects of the same intervention for people from different strata. By combining CRFs with the rich structure of ECLS-K: 2011, future research can move from identifying risk indicators to evaluating the real-world influences of policy levers, such as screening protocols, classroom accommodations, and family outreach, on narrowing achievement gaps.

In conclusion, we observe the noticeable difference between the two groups, indicating that it is necessary to do early and multi-tiered screening, allocate more resource that prioritizes foundational numeracy, literacy, and self-regulation; and culturally responsive practices in under-resourced schools. Adopting advanced causal methods will be essential for designing and assessing interventions that not only reduce current inequity but also deliver scalable solutions to support every child's learning journey.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] National Center for Education Statistics (2016) Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K:2011). U.S. Department of Education, Institute of Education Sciences.

[2] Holland, P.W. (1986) Statistics and Causal Inference. *Journal of the American Statistical Association*, **81**, 945-960. https://doi.org/10.1080/01621459.1986.10478354

[3] Pearl, J. (2009) Causality: Models, Reasoning, and Inference. 2nd Edition, Cambridge University Press. https://doi.org/10.1017/cbo9780511803161

[4] Rubin, D.B. (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66**, 688-701.

https://doi.org/10.1037/h0037350

[5] Lalonde, R.J. (1986) Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, **76**, 604-620.

[6] Rosenbaum, P.R. and Rubin, D.B. (1983) The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41-55. https://doi.org/10.1093/biomet/70.1.41

[7] Dehejia, R.H. and Wahba, S. (1999) Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, **94**, 1053-1062. https://doi.org/10.1080/01621459.1999.10473858

[8] Hirano, K., Imbens, G.W. and Ridder, G. (2003) Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, **71**, 1161-1189. https://doi.org/10.1111/1468-0262.00442

[9] Austin, P.C. (2011) An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, **46**, 399-424. https://doi.org/10.1080/00273171.2011.568786

[10] Athey, S. and Imbens, G. (2016) Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences*, **113**, 7353-7360. https://doi.org/10.1073/pnas.1510489113

[11] Varian, H.R. (2016) Causal Inference in Economics and Marketing. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 7310-7315. https://doi.org/10.1073/pnas.1510479113

[12] Keller, B. and Tipton, E. (2016) Propensity Score Analysis in R: A Software Review. *Journal of Educational and Behavioral Statistics*, **41**, 326-348. https://doi.org/10.3102/1076998616631744

[13] McConnell, K.J. and Lindner, S. (2019) Estimating Treatment Effects with Machine Learning. *Health Services Research*, **54**, 1273-1282. https://doi.org/10.1111/1475-6773.13212

[14] Chipman, H.A., George, E.I. and McCulloch, R.E. (2010) BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, **4**, 266-298. https://doi.org/10.1214/09-aoas285

[15] Athey, S. and Wager, S. (2019) Estimating Treatment Effects with Causal Forests: An Application. *Observational Studies*, **5**, 37-51. https://doi.org/10.1353/obs.2019.0001

[16] Chesnaye, N.C., Stel, V.S., Tripepi, G., Dekker, F.W., Fu, E.L., Zoccali, C., *et al.* (2021) An Introduction to Inverse Probability of Treatment Weighting in Observational Research. *Clinical Kidney Journal*, **15**, 14-20. https://doi.org/10.1093/ckj/sfab158

[17] Morgan, P.L., Frisco, M.L., Farkas, G. and Hibel, J. (2008) A Propensity Score Matching Analysis of the Effects of Special Education Services. *The Journal of Special Education*, **43**, 236-254. https://doi.org/10.1177/0022466908323007

[18] Harry, B. and Klingner, J. (2006) Why Are So Many Minority Students in Special Education? Teachers College Press.