Earlier Preschool Education for Better Development? Evidence from Vietnam

An Nguyen, Stéphane Goutte

Earlier Preschool Education for Better Development? Evidence from Vietnam

An NGUYEN* Stéphane GOUTTE†

May 19, 2022

Abstract

Using longitudinal Young Lives survey in Vietnam, this paper analyzes the effect of preschool starting age on skills, educational attainment, and health of poor children. Implementing the double machine learning estimator, we find short and medium-term positive effects of starting preschool younger on child cognitive skills but little evidence on child non-cognitive skills. Starting preschool younger significantly increases the highest grade achieved and creates a behavioral response of increasing subsequent parental investment in education. In contrast, we uncover the short-term negative impact of starting preschool younger with measures of child health, namely weight-for-age and BMI-for-age. The findings imply that expanding access to preschool and improving the quality of preschool education are important to achieve comprehensive child development.

JEL Classification: J24, O15, I26
Keywords: Early childhood education, Preschool education, Double Machine Learning

1 Introduction

Does it matter at what age disadvantaged children attend preschool? Although an accumulating body of evidence shows the beneficial effects of preschool education on disadvantaged children (Berlinski et al., 2008; Figlio and Roth, 2009;

*University Paris-Saclay, UMI SOURCE, IRD, UVSQ, France & IPAG Business School; ngth.hoai.an@gmail.com
†University Paris-Saclay, UMI SOURCE, IRD, UVSQ, France & Internationale School, Vietnam National University, Hanoi, Vietnam; stephane.goutte@uvsq.fr

Fredriksson et al., 2010; Felfe and Huber, 2016), none of the studies reviewed appears to assess directly the effect of preschool starting age on child outcomes.

Early childhood experiences have persisted and profound impact on brain development—affecting the accumulation of life-relevant skills and, ultimately, well-being (Sapolsky, 2004; Knudsen et al., 2006). Adverse early childhood environments raise serious concerns about the accumulation of these skills and eventually the life prospects of disadvantaged children. Recent studies prove the beneficial effect of high-quality early childhood education programs on the development of disadvantaged children by enriching their learning and nurturing environments. Theoretical evidence proposed by Cunha et al. (2010) also suggests that early childhood education can have a long-term impact on skills by directly improving current skills, which will beget later skills (i.e., self-productive) and increasing the productivity of later investment (i.e., dynamic complementarity). Based on these shreds of evidence, we might expect that early preschool enrollment would benefit disadvantaged children by preventing the negative effect of early life conditions and shaping the skills that promote their later life outcomes.

The question is specifically relevant to a country such as Vietnam, which already has universal preschool take-up for children one year before primary school. The proven benefit of early preschool enrollment will motivate countries like Vietnam to promote or even to require children from disadvantaged households, local poor areas, and ethnic minorities to enter preschool early to tackle poverty and inequality in the long run.

We address this question by analyzing the effect of preschool starting age on the development of poor children in Vietnam. In exploring this relationship, the contribution of our study is three folds. First, instead of focusing on pure preschool enrollment as most studies do, we directly tackle the effect of

preschool starting age on a wide range of outcomes, including cognitive skills, non-cognitive skills, health, and educational attainment over a 10-year development period of children. This broad coverage of outcomes over a long period makes it possible to have a complete picture of the impact until adolescence. Second, we also contribute to the literature by assessing the potential pathway through which early preschool enrollment improves child outcomes suggested by theory. Understanding the underlying mechanisms can help design and improve the early childhood education policy for disadvantaged children to benefit these children and the entire society. Finally, the paper contributes to the causal machine learning literature by applying the double machine learning estimator (Chernozhukov et al., 2018) empirically for average treatment effects. The method depending on conditional independence assumption such as propensity score matching requires identifying a set of control covariates enter the analysis and the functional form. Many times the functional form is not clear, and a large number of control variables is needed. The recently introduced framework of double machine learning (Chernozhukov et al., 2018) allows the use of many control variables and their flexible functional form in an objective and data-driven way to make the conditional independence assumption (CIA) plausible.

We use the Young Lives Survey, which aims to document child poverty in Vietnam. It yields a wealth of measurement of child skills and information of the household, community, and parental investments in children. The highly dimensional data makes use of the conditional independence assumption likely plausible by providing an enormous set of control variables. The treatment variable is the preschool starting age of children. Outcomes include cognitive skills measured by mathematics and language; non-cognitive skills captured by self-efficacy, self-esteem, parent-child relations, and peer relations; and health captured by height-for-age, weight-for-age, and BMI-for-age z-scores. We also

look at other outcomes comprising education attainment and nutritional statuses. These measures cover a wide range of aspects of child development that can be positively affected by early preschool enrollment. Finally, to analyze a potential channel through which early preschool enrollment affects child development, we assess the impact on parental investment in education proxied by the number of hours of extra classes.

Exploiting high-dimensional data and the advance of machine learning, we reach the following findings. First, we uncover short- and medium-term favorable effects of early preschool enrollment on child cognitive skills and educational attainment but little evidence of the impact on non-cognitive skills. Specifically, at 15, a one-year younger preschool starting age raises child language skills by 0.091 standard deviations and increases the highest grade achieved by 0.047 year. We find that the favorable impacts could be, at least in part, explained by parental investment behavior (proxied by the number of hours of extra classes children attended) as early preschool enrollment is associated with an increase in the number of hours of extra classes in all ages observed in the survey. Second, we analyze the impact on child health using anthropometric scores. We uncover the unfavorable short-term effect of early preschool enrollment on weight-for-age and BMI-for-age z-score. A one-year earlier preschool enrollment decreases weight-for-age z-score and BMI-for-age z-score by 0.056 and 0.092 standard deviation at age 8, respectively. This finding was unexpected and suggests that the quantity of nutritious value of preschool feeding might be less than home feeding. Interestingly, there is a negative association between preschool starting age and the probability of being thin at age 15, suggesting a link between health outcomes and educational attainment.

Although previous literature proves that the positive effect of preschool enrollment is higher for disadvantaged children, our study casts doubt on the qual-

ity of school feeding for poor children in Vietnam. While preschool in Vietnam might enrich the cognitive learnings of poor children compared to childcare alternatives or home environments, we argue that improving the quality of school meals and non-cognitive skill stimulation is essential to achieve comprehensive child development.

The paper unfolds as follows. Section 2 presents the relevant literature. Section 3 describes the data. Section 4 outlines the methodology. Section 5 presents our empirical results, and section 6 concludes.

2 Literature Review

Whereas the literature has provided evidence on the effectiveness of universal preschool education on outcomes of disadvantaged children, they only look at the effect of pure preschool attendance without considering variation in age when children enter preschool. A dearth of non-cognitive skills, long-term outcomes, and developing contexts also limits this current stock of evidence. There is little understanding about the medium- or long-run effects of preschool attendance for health, non-cognitive skills, as well as parental investment. Even though research has remained mainly confined to developed countries, the mechanisms by which it has generated the impacts have not been fully discovered. Analyses of preschool education in developing country contexts are still very limited and inconclusive. Assessing the role of preschool education on disadvantaged child development is especially important in these contexts, where expanding access to preschool provision has been one of the main education policies to addressing inequality, breaking the cycle of poverty, and improving outcomes later in life.

Within the context of developing countries, the recent study of Shafiq et al. (2018) attempts to assess the relationship of participation in early childhood education (ECE) and various long-term outcomes in 12 low- and middle-income

countries. They find positive and statistically significant associations between ECE participation and post-ECE educational attainment but fewer cases of positive associations between ECE and long-term socio-emotional outcomes. They also find mixed evidence of ECE and labor market outcomes, with positive associations for skill use but weak associations with earnings. Other studies of the impact of preschool education on long-term outcomes mainly have been confined in developed countries and on outputs such as school enrollment or test scores. Fredriksson et al. (2010) analyze the effects of preschool attendance on test scores at age 13 for four cohorts of children born in Sweden. Controlling for general cohort-specific trends, they find that preschool attendance significantly reduces the language score disparity between children of immigrants and those of native-born parents but does not affect the gap in scores on another test or academic secondary school degree completion. Berlinski et al. (2008) using data on children aged 7-15 in Uruguay and controlling for family fixed effects, find that preschool attendance is positively associated with school attendance and level of education with substantial effect for children with low-educated parents or living outside the capital city. Exploiting variation in child care prices as an instrument for preschool participation, Black et al. (2014) uncover a positive effect government-subsidized preschool on children's future national exam grades at age 16 in Norway, with the most significant impacts for children from low-income families. Hazarika and Viren (2013) study effects on adolescents in India using presence in the village of a center or a school offering preschool classes as an instrumental variable. They find that preschool participation increases school enrollment and also speeds grade progression conditional on enrollment. Drange et al. (2016) using a difference-in-difference approach to study the long-run effect on schooling of mandating kindergarten at age 5–6 in Norway, find no effect both overall, across sub-samples, and over the grading distribution.

Dhuey (2011) using similar strategy, detects the negative effect of kindergarten expansions on grade retention among Hispanic children, non-English speakers, children of immigrants, and children from low socioeconomic status households in the US.

The impact of universal preschool education on measurements of child skills (e.g. cognitive or non-cognitive capabilities) has only been assessed mainly short-term outcomes in developed countries, especially in the US. Using the birthday cut-off for enrollment to construct RDD, (Gormley Jr et al., 2005; Gormley et al., 2008) find positive effects of Oklahoma's universal pre-kindergarten program on children's literacy and math, as well as socio-emotional development, with the largest benefits for disadvantaged children. Fitzpatrick (2008) uses a difference-in-difference method to show that Georgia's universal pre-kindergarten program increased the reading and math scores of disadvantaged children and those in small towns or rural areas while reducing the probability of grade retention. Employing the presence or absence of a public pre-kindergarten program in the student's local zoned elementary school as an instrument, Figlio and Roth (2009) document that attending public pre-kindergarten school in Florida reduces behaviour problems and suspension or grade retention rates in short-term with the most significant benefits from children from families with low levels of education or disadvantaged neighborhoods. Applying the same approach with individual distance to the nearest preschool facility as intrumental variable, Felfe and Huber (2016) found significant short-term gains in terms of children's literacy. Preschool attendance also increases the prevalence of vaccinations but does not affect other observed health outcomes for Roma children in Europe.

When assessing the impact of preschool education, the fundamental identification problem is that preschool attendance is likely to be endogenous. The

current literature uses different approaches to tackle this problem. The rigorous research by far used the quasi-experimental method, their methodological strategies, yet have some crucial drawbacks. For example, several research exploiting regional and time variations by using the difference-in-difference framework (Fitzpatrick, 2008; Drange et al., 2016) is inherent to a crucial drawback that if there are any differences in trends of unobserved characteristics between treatment and control group, the difference-in-difference estimates capture the differences in trend rather than the effect of the programs. Other literature (Gormley Jr et al., 2005) exploited the nature of the data in the US to shape the regression discontinuity design on age eligibility. This strategy faces the problem of an insufficient number of children to identify the impact in the cut-off's vicinity and of selective attrition from the sample (Elango et al., 2016).

Our paper aims to address these concerns and contribute to the literature in three aspects. First, instead of focusing on pure preschool enrollment, we focus on the impact of preschool starting age on a broad set of outcomes representing the development of poor children over ten years. This applies to a country like Vietnam where already has universal preschool enrollment among five-year-old children. Second, we assess the relationship between preschool starting age and child development using a new empirical approach compared to the studies mentioned previously. In particular, we exploit the double machine learning approach dealing with a high-dimensional set of control variables while prior studies adopt the quasi-experiment method as the identification strategy. The richness of our longitudinal data allows us to construct numerous potential control variables. These variables controlled in flexible functional forms make the conditional independent assumption workable. Finally, we contribute to the literature by testing a potential channel through which preschool starting age improves child outcomes.

3 Data, Treatment Variable and Outcomes

3.1 Data

The data used here are the Young Lives Survey of Younger Cohort, which aims to document child poverty. The Young Lives survey uses quantitative and qualitative data collection methods, tracking children 2,000 children in five rounds in 2002, 2006, 2009, 2013, 2016, which correspond to the age of 1, 5, 8, 12, 15 of the Younger Cohort. The Young Lives data is intended to gather information on diversified life aspects of children, their caregivers, and families. The survey has collected comprehensive information on child characteristics, families, caregiver/parent characteristics and resources, their preferences and feelings, as well as school and community factors. Moreover, the survey provides specific information for measuring both children cognitive and non-cognitive skills, allowing us to study children development.

Although the Young Lives household survey sampling is not designed to represent the Vietnamese population, it considers the diversity of children in the country regarding various attributes and experiences. These surveys use a sentinel-site sampling design, including a complex consultative process with a pro-poor selection rule based on each commune's poverty ranking. Overall, the data consists of 20 purposely selected sites representing diversity and pro-poor bias in five provinces: Lao Cai, Hung Yen, Da Nang, Phu Yen, and Ben Tre. Each province contains four sites, and each site comprises one or two communes, in total, with 36 communes. At the site level, children have selected randomly in 2001 such that the data represents the birth cohort at each site.

3.2 Preschool starting age

There are three types of institutions for early childhood education in Vietnam, depending on the child's age. From three months to three years old, children can attend nurseries/creches. They go to kindergartens from three to six years old. Preschools in Vietnam incorporate the functions of these two schools and welcome infants from the age of 3 months up to the age of 6 years (Vu, 2021). From the age of 3, preschool functions as an educational facility that promotes developmental learning and growth and prepares children for primary education. As a result, this study only considers preschool starting age from three years old. This exclusion also makes preschool in Vietnam similar to that of other countries.

The Young Lives Survey provides us with a wide variety of child characteristics. The primary explanatory variable is the child's preschool starting age, collected in the second round when the child is five years old. We exclude children who did not attend preschool or started preschool before the age of three years, accounting for 11% and 0.7% of the total number of young cohort children in the survey, respectively. Table 1 reports the summary statistics of preschool starting age. The average years of preschool starting age are approximately 3.72 years old. Children start preschool latest at 5.75 years old.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Preschool starting age	1,763	3.723	0.757	3.000	3.000	4.167	5.750

Table 1. Summary statistics of preschool starting age variable

3.3 Child skills and education

We cover two dimensions of child cognitive skills, which are math skills and language skills. They are measured by mathematic tests (round 3, 4, and 5), Peabody Picture Vocabulary test (round 3,4 and 5), Early Grade Reading Assessment (round 3), and Reading Comprehension test (round 4 and 5). Based on the mathematics test and the combination of vocabulary and reading tests, we construct measures of mathematics and language skills using the Rasch model. The Rash model is a one-parameter of the Item Response Theory model using logistic regression analyses where item responses are predicted by the difference between the item difficulty and the child's ability. The equation for the Rasch model is:

$$P(\theta_j, b_i) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}},\ i = 1, 2, 3, ..., n$$

P is the probability that a child j with ability θ_j get the right item i. b_i is the item difficulty. n is the number of items in the test. θ_j is child's ability parameter.

We apply marginal maximum likelihood estimation to estimate item parameters and person parameters (Rasch score) and then construct θ score for each child.

The non-cognitive skills are measured by self-efficacy, self-esteem, parent-child relations, and peer relations in round 4 and 5 (Tran, 2017). We use factor analysis (Yong et al., 2013) to compute the non-cognitive skills based on respondent's degree of agreement or disagreement with some statements. Agreement ranges from "strong agreement" to "strongly agreement" with a 4-point scale; all the items are recorded to reflect positive statements. The number of items in the questionnaire related to self-efficacy, self-esteem, parent-child relations, and peer relations is ten, eight, nine, and eight, respectively. The simple math-

ematical model for factor analysis is:

$$X_j = a_{j1}F + e_j,\ j = 1, 2, ..., p$$

p denotes the number of items in the group $(X_1, X_2 \ldots, X_p)$. The factor loading a_{j1} is the factor loading of j^{th} variable on the factor F, and e_j is a residual variate. We assume that a group of items only represents one underlying non-cognitive skill dimension called factor. Since the data can violate the assumption of multivariate normality, we use the Principal Axis Factor for factor extraction and oblique rotation for rotation method, as suggested by Yong, & Pearce, 2013. We run four factor analysis for four measures of non-cognitive skills. Depending on the result, we only choose significant items having the value of factor loading greater than 0.4. We end up with nine, six, eight, and eight items in round 3 and six, six, eight, and seven items in round 4 to construct final measures of self-efficacy, self-esteem, parent-child relations, and peer relations, respectively. We also assess the impact on educational attainment using the highest grade achieved.

We explore one potential mechanism through which preschool starting age can affect child development by assessing its impact on parental investment in education proxied by the number of hours of extra classes. As suggested by theoretical evidence, early investment can promote subsequent investment and increase its return.

3.4 Child health

We measure child health with child's anthropometry, including height-for-age, weight-for-age and BMI-for-age z-score. In the Young Lives data, z-scores were estimated using World Health Organization (WHO) reference tables and software. All anthropometric scores adjusted by child's age and sex. Since the WHO

reference tables apply only to children of certain ages, the weight-for-age z score is computed for the Younger Cohort only up in Round 3. We also consider the nutritional statuses built up based on the child's anthropometry. The survey used the WHO table (2007) to construct Stunting, Underweight and Thinness based on height-for-age, weight-for-age and BMI-for-age z-scores, respectively. We observe three dummy variables taking the value of one if the corresponding z-score is below the tables' cut-off of -2, and zero otherwise. The summary statistics of all outcomes variables are shown in table appendix A

4 Empirical Framework

4.1 Identification

We estimate the effect of preschool starting age on child skills, long-term health and educational attainment. Our empirical strategy is motivated by a new double machine learning estimator for average treatment effects and the availability of longitudinal rich data of poor children in Vietnam. The double machine learning method and the advance of machine learning in solving prediction problems provide a promising way to draw causal inferences from observational data. The fundamental problem of confounding is that preschool attendance is likely endogenous since the preschool investment decision of parents reflects several factors such as their taste, resources, and information about the current evolution of their child skills that may be correlated with child skills, education and health.

To eliminate confounding, we rely on the existence of a vector of covariates X_i such that the following two assumptions are fulfilled:

1. Conditional independence: The treatment status D_i is as good as randomly assigned and not correlated with potential outcome Y_i^t conditional

on covariates X_i

$$Y_i^t \perp\!\!\!\perp D_i | X_i$$

2. Common support: Any unit needs to have a non-zero probability to receive each of the treatments:

$$P[D_i = t | X_i = x] > 0$$

In this case, the CIA requires observing all variables that influence the parent's decision to send children to preschool and the outcomes of interest simultaneously.

Following Cunha et al. (2010), inherited traits and prenatal investments affect skills at birth which are correlated with investments in early childhood and child outcomes. We control this concern by considering the child's characteristics, including child's age, child's sex, ethnic minority, number of brothers and sisters in the household, and child health using child's anthropometry at age 1, parental background and parental education. Moreover, household and community environments have an impact on both child investment and child development. We consider the household environment by controlling for household socioeconomic characteristics and the sentinel site environment by controlling for sentinel site id.

Parent assesses the cost of preschool depending on parental labour supply and other child care options. A working caregiver is more likely to search for other support. We control for the caregiver's working activity and the support caregiver can receive inside and outside of the household. The support can be childcare directly, in particular from household members, not household members or creche. She can also receive general support from groups for which she is a member or any other sources for which she can look. All these supports

represent social capital that can have an impact on the relative cost of preschool education.

Preschool investment decisions of parents also reflect parental taste, parental altruism towards the child and parental expectation. We include the caregiver's attitude and perception and child care history to control for these factors.

All the control variables are measured in the first round except for the caregiver's attitude and perception. We assume that the caregiver's attitude and perception are not updated until the child goes to primary school. In total, we extract 50 potential control variables from the Young Lives survey.

4.2 Estimation

Double machine learning approach (Chernozhukov et al., 2018) estimates outcome and treatment by machine learning algorithms and combines the estimates to derive average treatment effect estimator. The data-adaptive machine learning method provides small-mistake estimates of nuisance parameters to de-bias estimates of the average treatment effect.

The equations models outcome and treatment as follows:

$$Y_i = \theta_0 D_i + g_0(X_i) + e_i \qquad , E[e_i|D_i, X_i] = 0 \tag{1}$$

$$D_i = m_0(X_i) + u_i \qquad , E[u_i|X_i] = 0 \tag{2}$$

The confounding factors X_i affect the treatment variable D_i via the propensity score, $m_0(X_i) := E[D_i|X_i]$, and the outcome variable via the function $g_0(X_i)$. e_i and u_i are stochastic errors. Given the conditional independence assumption, θ_0 is interpreted as a structural or causal parameter that can be inferred from equation (1). The double/debiased ML estimator $\hat{\theta}_0$ solves the following condition:

$$E[\psi(W, \theta_0; \eta_0] = 0 \tag{3}$$

Where $\psi(W;\theta,\eta) = (Y - D\theta - g(x)(D - m(x))$ is score function, $W = (Y, D, X)$, θ_0 is the parameter of interest, and $\eta = (g, m)$ denotes nuisance functions with population value $\eta_0 = (g_0, m_0)$. The score ψ satisfies the *Neyman orthogonality condition* $\partial_\eta E\psi(W;\theta_0,\eta)|_{\eta=\eta_0} = 0$ which ensures that the condition (3) used to identify and estimate θ_0 is insensitive to small pertubations of the nuisance function η around η_0 and to the replacement of η_0 with $\hat{\eta}_0$ for high-quality inference. Therefore, the estimation $\hat{\theta}_0$ of θ_0 is constructed base on the score ψ which replace the nuisance functions η_0 with $\hat{\eta}_0$ produced by data-adaptive machine learning method and on the moment conditions (3).

This estimator is semi-parametric efficient, under weak conditions, due to an extra step of sample splitting. Since using one sample to fit the nuisance parameter, obtain prediction and construct the estimate of θ_0 causes overfitting and thus bias, sample splitting would help to avoid overfitting and thus reduce bias. Chernozhukov et al. (2018) propose randomly split the sample to $k-$folds equal partitions $(1, 2, \ldots, N)$. For each set I_k, let I_k^c denote all observation indices that are not in I_k. The nuisance parameters are estimated in the auxiliary sample I_k^c: $\hat{\eta}_{0,k} = \hat{\eta}_{0,k}\big((W_i)_{i\in I_k^c}\big)$. Then we use this estimation to construct the estimation $\hat{\theta}_0$ as the solution to the equation:

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi(W_i;\hat{\theta}_0,\hat{\eta}_{0,k}) = 0 \tag{4}$$

The asymptotic variance of $\hat{\theta}_0$ is estimated by:

$$\hat{\sigma}^2 = \hat{J}_0^{-2}\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}[\psi(W_i;\hat{\theta}_0,\hat{\eta}_{0,k})]^2$$

$$\hat{J}_0 = \frac{1}{N}\sum_{k=1}^{K}\sum_{i\in I_k}\psi_a(W_i;\hat{\eta}_{0,k})$$

$$\text{where } \psi_a(W;\eta) = -(D - m(X))(D - m(X))$$

The specific sample splitting creates an additional source of variation, having impact on estimation results asymptotically in finite samples (Chernozhukov et al., 2018). They suggest further adaptation by re-partitioning the data multiple times and then aggregate it by the median of the resulting estimates as the final average treatment effect. The estimated standard errors are also corrected to capture the spread of the estimates.

We implement the double machine learning approach of Chernozhukov et al. (2018) applying the `DoubleML` package (c, 2021) in `R`. We use eXtreme Gradient Boosting - XGBoost by Chen and Guestrin (2016) to fit the nuisance parameters. XGBoost is a gradient boosted decision-tree-based ensemble Machine Learning algorithm that focuses on computational speed and model performance[1]. This model does not require the form of nuisance parameters to be specified and automatically handle missing data, implicitly conduct feature selection and characteristics. XGBoost is fast and dominates structured or tabular datasets on classification and regression predictive modelling problems [2]. We fit XGBoost to estimate each nuisance function with parameters chosen by internal tunning using five-fold cross-validation.

5 Results

We report estimates obtained using five-fold cross-fitting. All results are based on taking 100 different sample splits. The results are summarized across the sample splits using the median method.

[1] Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

[2] Since its introduction, this algorithm has been credited with winning numerous Kaggle competitions

5.1 Impact on skills, educational attainment and parental investment

Panel A and B of table 2 shows the results for the effect of preschool starting age on child cognitive skills. All skill variables are standardized to have zero mean and variance one. The double machine learning estimates show a negative effect of preschool starting age on mathematics and language skills at age of 8. The estimates imply that starting preschool one year younger increases the mathematics score by about 0.09 standard deviations and language score by about 0.06 standard deviations. At age of 12 and 15, attending preschool one year earlier is only significantly associated with an increase in child's language skills, in which the estimates are almost 0.07 and 0.09 standard deviations, respectively. In particular, the magnitude of these estimates tends to increase over time. Under the light of the technology of skill formation proposed by Cunha et al. (2010), these findings support the idea of *self-productivity* for language skills in which "skills beget skills". However, we do not find enough evidence of *self-productivity* for mathematics skills.

We find little evidence on the effect on non-cognitive skills. As shown in panel B of table 2, early preschool enrollment is only significantly associated with self-esteem at age 12, but the effect disappears when children reach age 15. The estimates of other dimensions of non-cognitive skills turn out not statistically significant. Overall, panel C reports the impact on educational attainment. Our estimates suggest that starting preschool one year younger raises the highest grade achieved by 0.034 year at age 8 and 0.47 year at age 15.

Our estimated impacts of starting preschool younger converge on the prior research about the impact of preschool participation reviewed earlier. In particular, Shafiq et al. (2018) show positive statistically significant and positive associations between early childhood education participation and long-term cognitive

skills and educational attainment in Vietnam. However, using different measurements of non-cognitive skills, they find mixed evidence on socio-emotional skills with both negative and positive impacts. Their data consist of urban adults from 20 to 64 years old in 2014, so this would not have been the same preschool program and underlying population with our study. However, an implication of the findings is the possibility that preschool education in Vietnam might focus on cognitive stimulation rather than enrich the learning of non-cognitive skills.

We explore an important channel through which preschool starting age affect child skills and child education. According to Cunha et al. (2010), early investment increases the stock of future skills that promote the productivity of future investment. We test the hypothesis that early preschool enrollment boosts the future parental investment in education, explaining the effect of preschool starting age on child's outcome. Estimates of preschool starting age on parental investment in education are reported in panel C of table 2. We use the number of hours of extra classes as a proxy for parental investment in education. Attending preschool earlier is accompanied by short- and medium-term increases in the number of hours children attend extra classes during an average week, and indeed, the magnitude of effect increases over time. Starting preschool one year younger increases the number of extra-class hours on an average week by 0.576, 0.582 and 0.673 at age 8, 12, and 15 years, respectively. The significant short and medium-term effects on parental investment in education can partly explain the positive effect of early preschool enrollment on educational attainment.

5.2 Impact on Health

We present the estimates of the effects of preschool starting age on child health in Table 2. Panel A presents results on height-for-age, weight-for-age, and BMI-for-age score. Panel B shows the impacts on nutritional statuses. We find a

	(1)	(2)	(3)
	Age 8	Age 12	Age 15
Panel A: Cognitive Skills			
Mathematics	-0.089^{***}	-0.034	-0.051
	(0.034)	(0.037)	(0.038)
Language	-0.061^{*}	-0.077^{**}	-0.091^{**}
	(0.033)	(0.036)	(0.037)
Panel B: Non-cognitive Skills			
Self-esteems		-0.090^{**}	-0.007
		(0.042)	(0.043)
Self-efficacy		-0.050	-0.049
		(0.043)	(0.043)
Parent-child Relation		-0.022	0.007
		(0.044)	(0.042)
Peer Relation		-0.069	-0.03
		(0.044)	(0.045)
Panel C: Educational Attainment			
Highest grade achieved	-0.034^{**}	-0.032	-0.047^{*}
	(0.017)	(0.025)	(0.025)
Panel C: Parental Investment			
Number of hours of extra-classes	-0.576^{***}	-0.582^{***}	-0.673^{**}
	(0.192)	(0.176)	(0.251)

Note: *p<0.1; **p<0.05; ***p<0.01. Outcome variables of Panel A and B are scaled to have mean 0 and variance 1. Estimated effects and standard errors from partially linear regression effect model based on orthogonal estimating equations. Results are based on 100 sample splits and the standard errors adjusted for variation across splits are provided in parentheses.

Table 2. Estimated effect of preschool starting age on child skills, education attainment and parental investment.

	(1)	(2)	(3)
	Age 8	Age 12	Age 15
Panel A: Child's anthropometry			
Height-for-age z-score	0.005	−0.011	−0.011
	(0.033)	(0.033)	(0.034)
Weight-for-age z-score	0.073**		
	(0.033)		
BMI-for-age z-score	0.089**	0.045	0.044
	(0.038)	(0.039)	(0.040)
Panel B: Nutritional statuses			
Stunting	0.000	−0.009	−0.010
	(0.014)	(0.014)	(0.012)
Underweight	−0.008		
	(0.015)		
Thinness	−0.015	−0.005	−0.026**
	(0.012)	(0.013)	(0.013)

Note: *p<0.1; **p<0.05; ***p<0.01. Outcome variables of Panel A are scaled to have mean 0 and variance 1. Estimated effects and standard errors from partially linear regression effect model based on orthogonal estimating equations. Results are based on 100 sample splits and the standard errors adjusted for variation across splits are provided in parentheses.

Table 3. Estimated effect of preschool starting age on child health.

short-term negative effect of starting preschool younger on weight-for-age and BMI-for-age z-score. Starting one year younger is associated with a decrease in weight-for-age and BMI-for-age z- score by 0.073 and 0.089 standard deviations. Given gender, BMI-for-age and weight-for-age reflect current and long-term health status, respectively (Le and Nguyen, 2020). It can therefore be suggestive that preschool starting age can harm the current health of young children. We do not find enough evidence for the long-term impact of preschool starting age on health. We also find no evidence on nutritional statuses except for being thin at age 15 years. Incorporated with the effect of education, this can support the link between educational attainment and health outcomes.

6 Conclusion

While most studies have shown the favorable effect of preschool enrollment on child outcomes, no study, to the best of our knowledge, investigate the impact of preschool starting age. We apply the double machine learning approach given the conditional independence assumption to investigate the effect of preschool starting age on development of poor children in Vietnam. Our results suggest that early preschool enrollment has significantly positive short- and medium-term effects on child cognitive skills and educational attainment. However, we find little evidence of the effect on non-cognitive skills. We find that the favorable effect of early preschool enrollment can be, at least, partly explained by an increase in parental investment in education. Specifically, earlier preschool enrollment promotes the number of hours of extra classes in all rounds of the survey.

We uncover the unfavorable short-term effect of early preschool starting age on two measures of child health: BMI-for-age and weight-for-age, while we find no effect on height-for-age, the probability of the child being stunted, and

underweight. Although starting preschool younger does appear to benefit child skills and education, the downsides for child health imply that the quantity and quality of school feeding remain low for poor children. Nonetheless, we find a positive effect on the probability of the child being thin when children reach the age of 15 years. Incorporating with the prior results, this might be suggestive of a link between health and education attainment.

The positive effect of younger preschool starting age provides a signal that "earlier is better" for the development of poor children. However, preschool education in Vietnam might only enrich cognitive learning without nurturing environments for non-cognitive skills and health. Our findings suggest that improving the quality of preschool education is as important as, if not more than, expanding access to preschool to achieve sustainable development goals.

This study confines to poor children in Vietnam. Therefore its results cannot be generalized for the whole society. Further research needs to be done to assess the impact across different groups of children, cohorts, and countries.

Appendix A Summary statistics of outcome variables

	(1)	(2)	(3)
	Age 8	Age 12	Age 15
		Mean	
		(SD)	
		[N]	
Panel A: Cognitive Skills			
Mathematics	301.142	0.048	0.025
	(14.463)	(0.890)	0.917
	[1,749]	[1,759]	[1,740]
Language	304.465	0.059	0.048
	(10.487)	(0.855)	0.904
	[1,654]	[1,759]	[1,740]
Panel B: Non-cognitive Skills			
Self-esteem		0.001	0.007
		(0.823)	0.795
		[1,705]	[1,695]
Self-efficacy		0.015	0.014
		(0.839)	0.850
		[1,705]	[1,701]
Panel C: Educational Attainment			

	(1)	(2)	(3)
	Age 8	Age 12	Age 15
		Mean	
		(SD)	
		[N]	
Highest grade achieved		5.766	8.703
		(0.689)	(0.609)
		[1,732]	[1,691]
Panel D: Parental Investment			
Number of hours of extra-classes	6.845	5.217	4.911
	(5.976)	(5.202)	(5.821)
	[1,752]	[1,722]	[1,734]
Panel E: Health			
Height-for-age z-score	−1.039	−0.994	−0.987
	(1.047)	(1.125)	0.904
	[1,748]	[1,713]	[1,739]
Weight-for-age z-score	−1.078		
	(1.318)		
	[1,744]		
BMI-for-age z-score	−0.645	−0.576	−0.530
	(1.424)	(1.286)	1.186
	[1,740]	[1,713]	[1,739]
Stunting	0.208	0.215	0.127
	(0.460)	(0.490)	0.375

	(1)	(2)	(3)
	Age 8	Age 12	Age 15
		Mean	
		(SD)	
		[N]	
	[1,748]	[1,713]	[1,738]
Underweight	0.271		
	(0.524)		
	[1,743]		
Thinness	0.132	0.154	0.116
	(0.388)	(0.415)	0.376
	[1,737]	[1,713]	[1,738]

Appendix B Summary statistics of control variables

Variable	Mean	N	Mean	St. Dev.
Age 5				
Child's sex (male = 1, female = 0)	Binary	1,763	0.512	0.500
Ethnic minorities (Ethnic minorities = 1, Kinh = 0)	Binary	1,763	0.121	0.327
BMI-for-age z-score at age 1	Continuous	1,758	−0.412	0.983
Height-for-age z-score at age 1	Continuous	1,758	−1.102	1.311

Variable	Mean	N	Mean	St. Dev.
Weight-for-age z-score at age 1	Continuous	1,758	−0.956	1.088
Father's age	Discrete	1,704	30.006	5.861
Father's education	Discrete	1,697	6.478	4.112
Mother's age	Discrete	1,759	27.092	5.680
Mother's education	Discrete	1,749	5.914	4.106
Average working days per week of caregiver	Discrete	1,648	1.257	0.538
Age order of siblings in the household	Categorical	1,761		
Index child is the eldest		5		
Index child is the youngest		925		
Index child has no siblings in the household		831		
Number of sisters in the household	Discrete	1,763	0.437	0.762
Number of brothers in the household	Discrete	1,763	0.332	0.516
Number of males aged 0-5	Discrete	1,763	0.138	0.365
Number of males aged 6-12	Discrete	1,763	0.208	0.460
Number of males aged 13-17	Discrete	1,763	0.079	0.292
Number of males aged 18-60	Discrete	1,763	1.268	0.694
Number of males aged 61+	Discrete	1,763	0.111	0.316
Number of females aged 0-5	Discrete	1,763	0.163	0.387
Number of females aged 6-12	Discrete	1,763	0.268	0.571
Number of females aged 13-17	Discrete	1,763	0.101	0.357
Number of females aged 18-60	Discrete	1,763	1.353	0.720

Variable	Mean	N	Mean	St. Dev.
Number of females aged 61+	Discrete	1,763	0.163	0.377
Wealth index	Continuous	1,763	0.461	0.217
Housing quality index	Continuous	1,763	0.584	0.303
Access to services index	Continuous	1,763	0.436	0.271
Consumer durables index	Continuous	1,763	0.364	0.208
Access to safe drinking water	Binary	1,763	0.102	0.302
Access to sanitation	1,763	0.524	0.50	
Access to electricity	1,763	0.873	0.333	
Access to adequate fuels for cooking	1,763	0.245	0.430	
Number of groups caregiver is a member of	Discrete	1,763	0.387	0.728
Number of groups from which the caregiver has received support	Discrete	1,763	0.254	0.591
Number of other sources from which the caregiver has received support	Discrete	1,763	2.776	1.151
Number of household members who support child financially	Discrete	1,762	2.462	1.052
Number of household members who care for the child	Discrete	1,762	2.886	1.431
Creche enrollment in first year	Binary	1,763	0.064	0.245
Child care help from outside household	Binary	1,763	1.657	0.475
Level of formal education caregiver want the child to complete	Discrete	1,738	15.100	1.914

Variable	Mean	N	Mean	St. Dev.
At age caregiver expects the child to start earning money to support household	Discrete	1,731	21.376	6.240
Caregiver expects the child lives close when grown-up	Categorical	1,752		
Not at all		351		
A little		227		
Somewhat		375		
Quite a lot		562		
A lot		237		
Caregiver expects the child provides financial assistance to younger siblings when grown-up	Categorical	1,735		
Not at all		134		
A little		240		
Somewhat		586		
Quite a lot		616		
A lot		159		
Caregiver expects the child helps you with housework when grown-up	Categorical	1,761		
Not at all		224		
A little		355		
Somewhat		600		
Quite a lot		477		

Variable	Mean	N	Mean	St. Dev.
A lot		105		
Caregiver expects the child provides financial assistance when grown-up	Categorical	1,757		
Not at all		245		
A little		229		
Somewhat		478		
Quite a lot		645		
A lot		160		
Caregiver expects the child cares for you she/he is old	Categorical	1,761		
Not at all		32		
A little		75		
Somewhat		134		
Quite a lot		895		
A lot		625		
Caregiver expects the child provides emotional support	Categorical	1,761		
Not at all		11		
A little		19		
Somewhat		64		
Quite a lot		727		
A lot		940		

Appendix C Items included in factor analysis of non-cognitive skills

Skills	Statements	
	Round 3	Round 4
Self-esteem	• Overall, I have a lot to be proud of • I can do things as well as most people • Other people think I am a good person • A lot of things about me are good • I do lots of important things • When I do something, I do it well	• I'm as good as most other people • Overall, I have a lot to be proud of • I can do things as well as most people • Other people think I am a good person • A lot of things about me are good • I do lots of important things • When I do something, I do it well
Self-efficacy	• When I am confronted with a problem, I can usually find several solutions	• When I am confronted with a problem, I can usually find several solutions

Skills	Statements	
	Round 3	Round 4
	• If I am in trouble, I can usually think of a solution.	• I am confident that I could deal efficiently with unexpected events.
	• I am confident that I could deal efficiently with unexpected events.	• It is easy for me to stick to my aims and accomplish my goals.
	• I can always manage to solve difficult problems if I try hard enough.	• I can remain calm when facing difficulties because I can rely on my coping abilities.
	• It is easy for me to stick to my aims and accomplish my goals.	• I can usually handle whatever comes my way.
	• I can remain calm when facing difficulties because I can rely on my coping abilities.	• Thanks to my resourcefulness, I know how to handle unforeseen situations.
	• I can usually handle whatever comes my way.	

Skills	Statements	
	Round 3	Round 4
	• Thanks to my resourcefulness, I know how to handle unforeseen situations. • I can solve most problems if I invest the necessary effort.	
Parent-child relation	• I like my parents • My parents like me. • My parents and I spend a lot of time together • I get along well with my parents • My parents understand me • If I have children of my own, I want to bring them up like my parents raised me • My parents are easy to talk to	• I like my parents • My parents like me. • My parents and I spend a lot of time together • I get along well with my parents • My parents understand me • If I have children of my own, I want to bring them up like my parents raised me • My parents are easy to talk to

Skills	Statements	
	Round 3	Round 4
Peer ralation	• My parents and I have a lot of fun together • I make friends easily • I am popular with kids of my own age • Most other kids like me • Other kids want me to be their friend • I have more friends than most other kids • I have lots of friends • I am easy to like • I get along with other kids easily	• My parents and I have a lot of fun together • I make friends easily • I am popular with kids of my own age • Most other kids like me • Other kids want me to be their friend • I have more friends than most other kids • I have lots of friends • I am easy to like

References

Samuel Berlinski, Sebastian Galiani, and Marco Manacorda. Giving children a better start: Preschool attendance and school-age profiles. *Journal of public Economics*, 92(5-6):1416–1440, 2008.

Sandra E Black, Paul J Devereux, Katrine V Løken, and Kjell G Salvanes. Care or cash? the effect of child care subsidies on student performance. *Review of Economics and Statistics*, 96(5):824–837, 2014.

c. DoubleML – An object-oriented implementation of double machine learning in R, 2021. arXiv:2103.09603 [stat.ML].

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Flavio Cunha, James J Heckman, and Susanne M Schennach. Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, 78(3): 883–931, 2010.

Elizabeth Dhuey. Who benefits from kindergarten? evidence from the introduction of state subsidization. *Educational Evaluation and Policy Analysis*, 33(1): 3–22, 2011.

Nina Drange, Tarjei Havnes, and Astrid MJ Sandsør. Kindergarten for all: Long run effects of a universal intervention. *Economics of Education Review*, 53:164–181, 2016.

Sneha Elango, Jorge Luis García, James J Heckman, and Andrés Hojman. *4. Early Childhood Education*. University of Chicago Press, 2016.

Christina Felfe and Martin Huber. Does preschool boost the development of minority children?: the case of roma children. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(2):475–502, 2016.

David Figlio and Jeffrey Roth. 1. the behavioral consequences of pre-kindergarten participation for disadvantaged youth. In *The Problems of Disadvantaged Youth*, pages 15–42. University of Chicago Press, 2009.

Maria D Fitzpatrick. Starting school at four: The effect of universal pre-kindergarten on children's academic achievement. *The BE Journal of Economic Analysis & Policy*, 8(1), 2008.

Peter Fredriksson, Caroline Hall, Elly-Ann Johansson, and Per Johansson. Do pre-school interventions further the integration of immigrants? evidence from sweden. 2010.

William T Gormley, Deborah Phillips, and Ted Gayer. Preschool programs can boost school readiness. *SCIENCE-NEW YORK THEN WASHINGTON-*, 320 (5884):1723, 2008.

William T Gormley Jr, Ted Gayer, Deborah Phillips, and Brittany Dawson. The effects of universal pre-k on cognitive development. *Developmental psychology*, 41(6):872, 2005.

Gautam Hazarika and Vejoya Viren. The effect of early childhood developmental program attendance on future school enrollment in rural north india. *Economics of Education Review*, 34:146–161, 2013.

Eric I Knudsen, James J Heckman, Judy L Cameron, and Jack P Shonkoff. Economic, neurobiological, and behavioral perspectives on building america's future workforce. *Proceedings of the national Academy of Sciences*, 103(27): 10155–10162, 2006.

Kien Le and My Nguyen. Shedding light on maternal education and child health in developing countries. *World Development*, 133:105005, 2020.

Robert M Sapolsky. Mothering style and methylation. *Nature neuroscience*, 7 (8):791–792, 2004.

M Najeeb Shafiq, Amanda Devercelli, and Alexandria Valerio. Are there long-term benefits from early childhood education in low-and middle-income countries? *Education Policy Analysis Archives*, 2018.

Ngo Thi Minh Tam Tran. *Skill formation and transition to productive livelihood in Vietnam*. PhD thesis, PSL Research University, 2017.

Thao Thi Vu. Early childhood education in vietnam, history, and development. *International Journal of Child Care and Education Policy*, 15(1):1–18, 2021.

An Gie Yong, Sean Pearce, et al. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, 9(2):79–94, 2013.