As discussed in the \nameref{sec:lit\_rev}, there are significant empirical challenges to isolating the causal effect of early childhood investments on later-life outcomes. One common challenge is selection bias, particularly in Indonesia, where kindergarten is not compulsory and in the 1990s kindergartens were almost entirely private.\footnote{ See ~\ref{fig:kinder\_numscatter}. Based on the PODES data, in 1990, 95.06% of kindergartens were private. That share grew to 96.63% in 2000.} Clearly, kindergarten included a significant opportunity cost and required a significant investment of resources by a family to pay tuition and physically transport their young child to a classroom when it’s not required by law. To meet this challenge, I employ three empirical strategies: Ordinary Least Squares (OLS) estimation, Mother Fixed-Effects, and then Instrumental Variable (IV) estimation.

\subsection{OLS}

My starting point is a basic OLS estimation of the effects of kindergarten, using the following model:

\begin{gather}

Y\_{if} = \beta\_0 + \beta\_1\text{KINDER}\_{if} + \beta\_2\mathbf{X}\_{if} + \beta\_3\mathbf{Z}\_f + \beta\_4\mathbf{K}\_f + \epsilon\_i

\end{gather}

where $Y\_{if}$ is the outcome variable for individual $i$ in family $f$, $\text{KINDER}\_{if}$ is a dummy variable for kindergarten attendance for individual $i$, $\mathbf{X}\_{if}$ is a vector of individual-level variables, $\mathbf{Z}\_f$ is a vector of household-level controls, and $\mathbf{K}\_f$ is a vector of community-level controls. Our coefficient of interest is $\beta\_1$, which represents the effect of kindergarten attendance on the educational outcome.

Critical to my OLS approach is that I control for the two forces driving household human capital decision-making: 1) household expenditure/resources and 2) the importance placed on education/human capital within the household.\footnote{ Examples of the former include the natural log of household per-capita expenditure for every available year, tracking household expenditure across the educational career of an individual (this exogenous variable ostensibly absorbs regional economic shocks as well) and whether a household has electricity, also captured over time. Examples of the latter include a mother's or household head's years of completed education, whether an older sibling attended kindergarten, and the number of visits to a doctor by each child. These controls are implemented in conjunction with more basic individual and community controls. For example, in my analyses I control for the number of times a mother has taken an individual child to the doctor in the last few months. This, when weighed against evaluations of the child's health, may be a proxy for that mother's willingness or desire to invest in the human capital of the child -- after all, ensuring health in early childhood has been demonstrated to greatly enhance human capital accumulation. \citep{Attanasio2020}}

As part of my OLS estimations, I also employ province fixed-effects--as it is clear that there is a heterogenous component of the relationship between kindergarten and education across provinces, as demonstrated in Figure ~\ref{kinder\_prov}.\footnote{ In \nameref{sec:results} I also investigate heterogeneity. This province panel variable is more general than the mother identification variable; therefore, while this fixed-effects is less powerful than mother fixed-effects, it also does not restrict my sample--as Figure ~\ref{fig:scatter\_switching} makes clear, there is sufficient variation in every province so that my `switching sample' (see below) is the same as my total sample).} Finally, I use cluster-robust standard errors for the OLS estimations; thus, I cluster standard errors at the kecamatan--the equivalent of a county in Indonesia--level, making my estimates heteroskedastic-robust. As I'll discuss in my results section and in greater detail in the appendix, I use a variety of post-estimation commands and bootstrap methods to ensure my results are robust and the assumptions of OLS hold up under scrutiny.

\subsection{Mother Fixed-Effects}

For as many variables I can control in the above OLS model, there are critical household variables that I simply don't have data for--in particular, related to unobservable characteristics such as the emphasis placed on education within a household. Therefore, there will always be omitted variables correlated with \textit{both} kindergarten attendance and the educational outcome variables in the OLS model.

This omitted variable bias motivates an oft-employed approach to examining the effects of early childhood interventions: family fixed-effects. \citep{Currie1993} \citep{Garces2002} \footnote{ The stronger instance of family fixed-effects is the twin studies I described in the \nameref{sec:lit\_rev}.} In particular, I employed mother fixed-effects, an approach that is powerful because it immediately controls for all variables constant within a household. Fixed-effects accomplishes this feat by comparing between only siblings; specifically, I employ mother fixed-effects, therefore only comparing between children of the same mother. The fixed-effects approach estimates the following equation:

\begin{gather}

Y\_{if} = \beta\_0 + \beta\_1\text{KINDER}\_{if} + \beta\_2\mathbf{X}\_{if} + \beta\_3\mathbf{M}\_{ft} + \mu\_f + \epsilon\_i

\end{gather}

Note that the variables constant within the family from equation (1) are replaced with a single fixed effects term, $\mu\_f$, which captures \textit{all} mother characteristics--including all those that were not observed as part of the IFLS. Fixed-effects, however, is not a panacea; it only controls for all \textit{time-invariant} characteristics of the mother, rooted in 1997--the initial point of my sample. Therefore, I need to incorporate a vector of time-varying mother/household characteristics, represented by the term $\mathbf{M}\_{ft}$. \footnote{ These characteristics include economic or health shocks to the mother/household, the household's economic status, the household's structure (i.e., whether it remains a two-parent household or its size), and community-level shocks such as natural disasters.} It is possible that in reality a mother could treat their children differently—accordingly I’ve attempted to control for some pre-kindergarten characteristics that could signal favoritism. Then my model would be biased, as the fixed-effects model assumes $\mu$, i.e. a mother’s influence on a child’s kindergarten attendance and educational attainment, is fixed for all children. For more detail on this phenomenon, see \nameref{sec:lit\_rev}. \citep{Garces2002}

To avoid the bias of unobserved mother favoritism, I include a variable representing the number of times a parent took a child to the hospital, with health controlled for. It's also possible there are spillover effects from one child attending kindergarten on successive children--something I attempt to control for by including a dummy variable for whether an older sibling had already attended kindergarten. Lastly, it's possible that a family has a different quantity of resources at its disposal when one child is of kindergarten age than when another child is--something I control for by employing controls for time-varying household characteristics captured by household expenditures. The latter is a particular concern, as fixed-effects controls for all unobserved \textit{time-invariant} characteristics. I alleviate this concern by controlling for age; unfortunately, due to the low frequency of IFLS waves, I don't have standard data across the ages of siblings.

Fixed-effects introduces another complication; in order to run a regression for a specific family $f$, there must be variation in the explanatory variable, introducing a two-step requirement to be included in the fixed-effects model specification. These requirements are that 1) each mother has more than one child in our sample and 2) there is variation among the mother's children in kindergarten attendance. This significantly restricts our sample--as seen in the diagnostic graph Figure ~\ref{fig:scatter\_switching}.\footnote{ 17.48\% of the sample are from households with exactly one child. Only 7\% of my total sample qualifies as part of a switching household according to these requirements--meaning there's an attrition rate of 93\% just for the fixed-effects sample specification.}

These requirements for my fixed-effects specifications thus distinguishes between two types of families already included in my sample: 1) ``switching" families and 2) ``non-switching families". ``Switching" families are included in the fixed-effects model and ``non-switching families" are not. Often scholars are concerned about selection bias into the ``switching" designation. I find the ``switching" sample has some puzzling characteristics when compared to the total sample. There's no significant positive selection bias and the basic relationship between kindergarten and educational attainment is the same (estimated using OLS and IV) across switching and non-switching. However, selection into kindergarten attendance appears to be a drastically different process--something I discuss in greater detail in \nameref{sec:kinder\_sel} and \nameref{sec:switching}.

\subsection{Instrumental Variable (IV) Estimation}

Each of my first two empirical approaches--OLS and Fixed-Effects--have shortcomings. First, OLS merely suggests an association between kindergarten attendance and educational outcomes. Critically, it suffers from omitted variable bias because of a host of intangible measures. Second, as powerful as fixed-effects is for controlling household factors, it introduces significant attrition, 93\%, because of its switching requirements. In the end, the residuals from my mother fixed-effects estimation are correlated with kindergarten attendance--suggesting endogeneity and that my results are biased.

This motivates me to use instrumental variable (IV) estimation methods to remedy the bias introduced by the endogenous regressor, kindergarten attendance. I develop two instruments, both taken from the Village Potential Surveys (PODES) of 1990 and 2000, which contains explicit data about the presence of kindergartens in villages--see \nameref{sec:podes} for more information. I aggregate population and kindergarten across each kabupaten and then merge based on the kabupaten an individual or household is registered under in the 1997 wave of the IFLS, creating a population-weighted average of kindergartens for each kecamatan. My first instrument consists of total (private + public) kindergartens per 10,000 people in each kecamatan in 1990 and my second instrument is the same measure in 2000. \footnote{ In \nameref{iv\_robust} I explore alternative instruments, such as comparing just private to just public kindergartens, taking an average of the 1990 and 2000 measures, and percent change in the presence of kindergartens.} While I use the total number of kindergartens, the data clearly reflects it is private kindergartens driving expansion in the total number of programs between 1990 and 2000, as seen in Figure ~\ref{fig:kinder\_numscatter}.

The employment of an instrument requires a robust defense--both intuitively and empirically. An instrument needs to be predictive of kindergarten attendance while having no impact on educational attainment when relevant factors are controlled for. \citep{Levitt2002} Intuitively, the presence of kindergartens satisfies these conditions. Whether there is a kindergarten in a community would have a significant, direct effect on the likelihood of attending kindergarten. And when one controls for the presence of elementary, junior high, and senior high school in a community as well as for an individual's kindergarten attendance--as I do in my OLS specifications--then the presence of kindergartens would likely not have an effect on educational outcomes.

These intuitions are confirmed empirically. The first-stage regression results (found in Table ~\ref{table:iv\_first\_stage}) with my full set of covariates and the instruments as regressors and kindergarten attendance as the outcome variable, confirm the strength of my selected instruments. The F-Statistic in each of my specifications confirms the strength of my instruments--in the fully specified model, (4), the F-Statistic 124.92 is well beyond the `rule-of-thumb' benchmark of 10. Regarding the validity of the instruments, the second component, see model (3) in Table ~\ref{table:full\_results}: neither instrument, in the fully-specified OLS model, has a statistically significant coefficient. I explore the robustness and validity of my instruments further in \nameref{sec:results}, \nameref{sec:iv\_robust}, and \nameref{sec:iv\_post}.

There are different estimators for IV estimation; I employ the Generalized Method of Moments (GMM) estimator. In cases of over-identification, GMM doesn't reduce multiple instruments into one matrix and therefore is a more efficient estimator. It also gives me greater flexibility in conducting over-identification tests.