\subsection{Indonesian Family Life Survey}

For this paper, I use data from the Indonesian Family Life Survey (IFLS), a longitudinal household and community survey fielded through 5 primary waves from 1993 to 2014 by the RAND corporation. The availability of the IFLS--a free, public dataset--makes Indonesia unique among developing countries for having a detailed longitudinal household survey, which is necessary to study the medium- and long-term effects of early childhood interventions.

The original wave of the IFLS, fielded in 1993 and 1994, consists of 7,200 households and is representative of about 83\% of Indonesia's population at the time. \citep{Serrato1995} In each of the successive waves, the survey has attempted to re-interview the same households, in order to create a panel dataset. Successive waves have also been expanded to incorporate ``split-offs", i.e., households began by children of IFLS3 households as they aged up and became adults. In the fifth wave of the IFLS, 16,204 households and 50,148 individuals were interviewed. \citep{Strauss2016}

I leverage the advantages of the IFLS5 in a very clear way, by constructing a dataset of variables observed when individuals were young children in 1997, and variables observed when individuals were adults in 2014. Additionally, I incorporate community-level characteristics included in the IFLS--most importantly, for example, the number of schools per 10,000 people in a village/community--and household data taken from either the first or second waves of the IFLS. I do incorporate some aspects of panel data--namely in the repeated observations of household-level data such as household per capita expenditure or size of household.\footnote{For example, my independent variable--a dummy variable of whether an individual attended kindergarten--does not require repeated observation, nor do the \textit{outcomes} I'm interested in, namely school/grade completion or years of education completed.}

\subsection{Sample}

Thus, in order to standardize my sample size across my model specifications, I have stringent parameters for my sample. I begin my sample with all children from the ages 3 to 9 \textit{individually} interviewed in the `child' book of the 1997 household survey. This allows me to create early childhood individual-level controls that are responsible for variation between children of the same mother. Next, I require that those individually-interviewed children are listed on at least one household roster in 1997; the household roster is a part of the household survey that links members of a household together--I use it to link children to 1) their mother and, by extension 2) their siblings. The parameter that children were individually-interviewed, however, is more stringent than the parameter that they are listed on at least one household roster (there being many more children listed on a household roster than individually interviewed).

In order to evaluate medium- and long-term outcomes, I need to match these early childhood observations with later-life observations; for this, I rely on the 5th--and most recent--wave of the survey, fielded in 2014. Thus, for my third parameter, I require that all individuals in my sample were interviewed \textit{individually} (this time as adults, a designation which occurs when an individual is 15 and above for the purposes of the IFLS) in 2014. My fourth and final parameter is that there are no missing observations for \textit{any} of the variables used in my non-fixed effects regression specification. This includes observations of variables from the intervening--the 3rd and 4th--waves of the survey, which I've included. Thus, when including variables from the intervening waves of the survey, I've carefully balanced the merits of the tradeoff inherent in their inclusion--greater explanatory power (perhaps) at the cost of a smaller sample size (for some, a \textit{significantly} smaller sample size). The attrition results from these parameters--first resulting from moving from 1997 to 2014, and second resulting from the dropping of individuals with missing observations for my variables of interest--are discussed below, in the next section focusing on attrition.

See Table ~\ref{table:summary\_table} for summary statistics of the variables of interest for my sample. Clearly, there's \textit{some} correlation between 1) urban status and 2) kindergarten attendance, with both higher socioeconomic indicators (particularly for the household from which an individual originated) and educational outcomes. Also of importance is that the gap between children who went to kindergarten and those who didn't is larger, for a whole variety of the variables, among rural children than it is for urban children. This suggests possible heterogeneity in the effects of kindergarten attendance along the urban/rural split.

\subsection{Attrition}

I address the problem of non-random attrition in greater detail in the appendix. Overall, I find that while attrition \textit{is} non-random, there is not a significant positive bias for either category of attrition. As seen in Table \ref{table:attrition\_table}, if there is any bias it is a \textit{negative} selection bias into our sample. For example, an individual's mother's years of completed education is the variable with the most significant positive effect on educational outcomes in my non-fixed effects specifications. The mean of this variable, for observations within my sample, is significantly lower than for both those observations 1) lost to attrition due to missing observations and 2) lost to attrition due to moving from 1997 to 2014. This suggests that, \textit{with regards to attrition}, my estimates may be \textit{under-estimates} of the effects of kindergarten rather than over-estimates. Nonetheless, attrition is non-random--I use inverse probability weights to adjust for non-random attrition. The inclusion of these weights did not alter my results and they are discussed in greater detail in the appendix. \citep{Baulch2011}