\def \data #1{/Users/danielposthumus/thesis\_independent\_study/work/writing/rough\_draft/data/#1}

In this section, I describe the data and sample I use for my research. I begin by discussing my two sources of data--the Indonesian Family Life Survey (IFLS) and Village Potential Statistics (PODES), before I proceed to describing my sample and its construction.

\subsection{Indonesian Family Life Survey}

My main source of data is the Indonesian Family Life Survey (IFLS), a longitudinal household and community survey fielded through five waves from 1993 to 2014 by the RAND corporation. I use it for the entirety of my individual- and household-level data, as well as for the majority of my community-level data. The availability of the IFLS--a free, public dataset--makes Indonesia relatively 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 (with high rates of success in recontacting across the years), in order to create panel data. In the fifth wave of the IFLS, fielded in 2014, 16,204 households and 50,148 individuals were interviewed. \citep{Strauss2016} I discuss how I merged data from the survey's different waves and ensuing attrition in greater detail below as well as in the appendix.

I use the IFLS to track children from when they were young in 1997 to when they are adults in 2014. Some of my covariates take the form of panel data--namely in the repeated observations of household-level variables such as household per capita expenditure and size of household, or community-level variables such as the number of elementary schools per 10,000 people in each kecamatan.\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{Village Potential Statistics}

As comprehensive as the IFLS is, its community survey only began to inquire after the presence of kindergartens in a community in 2014: therefore, in order to construct my instrument--the number of kindergartens per 10,000 individuals in each kecamatan, I needed to supplement the IFLS with another data source.

The Village Potential Statistics (PODES) dataset is immensely powerful; PODES data is published every year, dating back to 1983. It's fielded by the Office of Statistics in Indonesia (BPS) and is publicly available only by purchase. As the price of purchasing PODES is dependent on the number of variables taken from the survey, I limit my use to the two variables I need: the presence of kindergartens and village population. I also limit the years of data to the ones I need--1990 and 2000. My research focuses on children who were kindergarten-aged during the 1990s; by collecting data that bookend the period of interest, I attempt to maximize data coverage.

PODES is a comprehensive and entirely representative of Indonesia, including approximately 65,000 villages (the exact number varies over time). Therefore, I sum the number of kindergatens (public, private, and total) as well as population by kecamatan and year (1990 and 2000) before finding the number of kindergartens per 10,000 individuals. I then merge this data with my IFLS data on each household's kecamatan in 1997.\footnote{ Since the PODES data is comprehensive, there's no attrition when merging the PODES data onto the IFLS data. This approach of merging IFLS data with selected variables from PODES is not novel and has been implemented in the literature. \citep{Shrestha2021}} A visualization of these variables can be found in Figure ~\ref{fig:kinder\_numscatter}.

\subsection{Sample Construction and Description}

The starting place for my sample is all children \textit{individually} interviewed in the `child' book of the 1997, or 2nd, wave of the IFLS. The `child book' consists of respondents less than 15 years old.\footnote{ This is Book 5 of IFLS2 -- individuals over 15 years old are interviewed in Books 3A and 3B.} 10,356 individuals completed this questionnaire. I then add an age restriction, keeping only the children who completed the survey and were between 3 and 9 years old in 1997. With this restriction, I was left with 5,258 individuals. I added an age restriction to ensure that only children who were of kindergarten age in the 1990s, per the years of my instrument, were included in the data.\footnote{ To exercise tighter control over age, I broke down individuals in my sample into 4 birth cohorts, those aged 3-4, 5-6, 7-8, 9-10. In terms of the number of households these individuals belonged to, I began with 5,171 unique households and, with the age restriction implemented, I was left with 3,488 unique households.}

Starting with children interviewed in 1997 constricts my sample; for comparison, 20,529 individuals were individually interviewed in the adult survey in the 3rd wave, which was much longer and detailed than the children's one.\footnote{ These individuals were from 7.538 unique households.} Yet, beginning here is necessary: I need to get a sense of the pre-kindergarten (or roughly pre-kindergarten) individual characteristics of the children, characteristics such as their health and how often parents took them to the doctor. I conduct robustness checks where I relax these specifications to test my hypotheses with fewer covariates and a larger sample size. \footnote{ This narrow sample specification is particularly problematic for my fixed-effects model, a challenge I discuss in greater detail in the Methods section and appendix.}

Another component of the household survey, in addition to individual interviews, is the household roster. The roster comprehensively documents each member of the household and their basic characteristics--such as their educational attainment and intra-household relationships. I use the rosters to link the 5,258 children from above to 1) their mother and, by extension 2) their siblings. \footnote{ There is no attrition in this step; children who are individually interviewed are all featured on a household roster, inclusion in the former being a much more stringent restriction than inclusion in the latter.} I then link the household roster to more detailed household survey data, such as whether a household has electricity, the size of the household, and the household's total expenditure. Finally, I merge these households with the IFLS's community survey, which contains observations of community variables such as population and the number of elementary schools in a community; by replacing communities that were not surveyed for the community survey with appropriate averages, I ensure there is no attrition with this step.\footnote{ Specifically, I'm working with 3 community-level variables: elementary schools per 10,000 people, junior high schools per 10,000 people, and senior high schools per 10,000 people. If an individual's community was not interviewed and thus these variables are missing, I replace them with the mean for their kecamatan, conditional on urban/rural status. If there is no data for the kecamatan, I replace their missing observations with the mean for the province, conditional on urban/rural status.}

Thus far, I have individual, household, and community data for individuals' pre-kindergarten and kindergarten-age; however, in order to evaluate medium- and long-term outcomes, I need to match these early childhood observations with later-life observations. First, I incorporate household and community data from 2000 and 2007 in order to absorb exogenous shocks to a child's situation as they grow. For example, I continue to track household expenditure levels and size over time. From 1997 to 2000, among the 5,258 children in my sample, only 80 were lost to attrition (1.5\% attrition rate). From 2000 to 2007, only 130 were lost to attrition (2.5\% attrition rate). These attrition rates are very low because I'm relying only on the very broad restriction that individuals appear on a household roster--which I use to merge household data onto individuals' observations.

Finally, I need individuals' educational outcomes, which are taken from the 2014 wave of the survey--when all of the individuals are over the age of 18. Therefore, for this restriction, I require that all individuals in my sample were interviewed \textit{individually} in 2014.\footnote{ This time as adults, a designation which occurs when an individual is 15 and above for the purposes of the IFLS.} From 2007 to 2014, 1,357 were lost to attrition (26.8\% attrition rate)--resulting in 3,691 individuals. For context, 36,391 individuals (aged 15+) were interviewed individually in 2014.

Lastly, I require that there are no missing observations for \textit{any} of the covariates used in my non-fixed effects regression specification. This leads to the attrition of 533 individuals, so that my final sample consists of 3,158 individuals (14.4\% attrition rate). Summary statistics of my sample for the variables of interest for my sample can be found in Table ~\ref{table:summary\_table}. More information on attrition can be found in the appendix and in Table ~\ref{table:attrition\_table}.

Clearly, there is some association between 1) kindergarten and 2) urban/rural status and both 1) educational outcomes as well as 2) household wealth or resources. It does appear that for most variables, the positive association between kindergarten and greater household resources or educational outcomes is stronger than that between urban residence and greater household resources or educational outcomes--a promising pattern. There's also clearly enough variation in my core variables--most of sample did not attend kindergarten (39\% did), and the average member of my sample completed 10.97 years of education, 1 year short of the completion of high school), while the median number of years of education completed is 12.

To give a sense of how representative the sample is of the whole Indonesian population, 95\% of the sample completed elementary school, 81\% completed junior high school, and 62\% completed senior high school. While historical data is difficult to match, pre-pandemic World Bank Data estimates that Indonesian students average 12.4 years of schooling and 37\% of 3-6 years were enrolled in preprimary education in 2018. \citep{Afkar2020} 2022 OECD data shows that 57.5\% of individuals 25-34 years old graduated senior high school, Thus, my sample appears to be broadly in-line with national summary statistics, although it may be that the sample is over-educated relative to the total Indonesian population.