Documentation for "Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia" by O. Attanasio, S. Cattan, E. Fitzsimmons, C. Meghir, and M. Rubio Codina

Data source

The raw data collected as part of the intervention is available on the UK Data Service website under Study number 851196 (https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=851196).

We appended municipality level data on conflict in 1995 from the El Panel Municipal del CEDE, which can be publicly accessed on the University of Andes website (https://datoscede.uniandes.edu.co/es/catalogo-de-microdata)

As part of the replication files we provide a dataset of cleaned variables. This dataset includes measures of child development, maternal skills, and investments at baseline and first follow up, as well as area characteristics that serve as instruments. The data is saved in **"data/dataset.dta"**.

Programs

a) Data preparation

- **"code/stata/globals.do"** create globals for directories and groups of variables.

- **"code/stata/prepare estimation sample.do"** selects the sample, standardizes the measures and create the bootstrapped samples. This do files creates the dataset "measures.dta", which contains all the variables that will be used in the estimation of the models.

b) Summary statistics and treatment impacts on raw measures

Tables 2, 3, 4 in the paper and Tables A1, C1, D1, D2 and D3 in the Online Appendix are obtained by working with raw measures included in the dataset "measures.dta". The folder "code/stata" includes do files that produces the results reported in each table. The name of the do file indicates which table it creates the results for.

c) Estimation of the model on the main sample

"code/R/MASTER.R" is the master R file that runs programs to estimate the measurement system, the factor scores and the investment and production functions reported in the paper and its appendix.

The first part of the "MASTER.R" program aims to estimate the measurement system. To do so, it calls a number of other programs, in the following order:

- **"EstimateMeasurementSystem.R"** : This program estimates the parameters of the measurement system on the true sample and/or bootstrapped samples.

- **"AssembleBootstrapFM.R"**: This program pulls together the estimates of the measurement system obtained from parallel R sessions in case this option was chosen to accelerate the running time. (See section (i) below for more details). It then computes the standard errors for the parameters of the measurement system and creates output tables for the parameters of the measurement system. These estimates are reported in Appendix Tables C.6 through C.10. This program needs to be run before factor scores can be estimated (even if the estimation of the factor scores was done in one R session).

Once the measurement system is estimated, the program calls the following two programs to compute the signal to noise ratio and graph the kernel densities of the latent factors:

- **"SignalToNoiseRatio.R"**: This program uses the estimates of the measurement system to compute the signal to noise ratio for each measure (Table 1 of the paper).

- **"KernelDensities.R"**: This program draws a large sample of latent factors from the estimated distribution, graphs their kernel densities and performs a Kolmogorov-Smirnov test of equality of distributions between treated and control groups on the bootstrapped samples (Figure 1).

The second part of the "MASTER.R" program aims to estimate the factor scores. To do so, it calls the following programs:

- **"EstimateFactorScores.R"**: This program estimates factor scores for every person in the sample using estimates of the measurement system as well as the observed measures.

- **"Assemble BoostrapFS.R"**: This program pulls together the factor score estimates obtained from parallel R sessions in case this option was chosen to accelerate the running time. This program needs to be run before the investment and production functions can be estimated (even if the estimation of the factor scores was done in one R session).

The third part of the "MASTER.R" program aims to estimate the investment functions and the production functions once the measurement system and factor scores are estimated. To do so, it calls a number of programs, which are all located in the subfolder "code/R/specsPF". Specifically:

- **"CobbDouglas_OLS.R"**: This program estimates the production function for cognitive and socio-emotional skills by OLS (column 1 of Table 6 and 7 in the paper).

- **"FirstStage_toyprice_foodprice_conflict.R"**: This program estimates the first stage equations and the reduced form equations using toy price, food price, and conflict as instrumental variables. The first stage results are reported in Table 5 (columns 1 and 2). The reduced form estimates are reported in Appendix Table E.1

- **"CobbDouglas_IV.R"**: This program uses the first stage and reduced form estimates generated by the previous program to estimate the structural parameters of the production function. The results of this specification are reported in Table 6, column 2 for cognitive skills and in Table 7, column 2 for socio-emotional skills.

- **“CobbDouglas_IV_notreatment.R”**: Same as the previous program, except that the production function does not include the treatment dummy (in other words, the treatment is used as an additional instrument). The results of this specification are reported in Table 6, column 3 for cognitive skills and in Table 7, column 4 for socio-emotional skills.

- **“CobbDouglas_IV_parsinv.R”**: Same as “CobbDouglas_IV.R”, except that the production function only includes on type of investment (material investment in the production function for cognitive skills and time investments in the production function for socio-emotional skills). The estimates of this specification of the production function for socio-emotional skills are reported in Table 7, column 3.

- **“CobbDouglas_IV_parsinv_notreatment.R”**: Same as “CobbDouglas_IV._parsinv.R”, with the exception that the production function does not include the treatment dummy (in other words, the treatment is used as an additional instrument). The estimates of this specification of the production function for socio-emotional skills are reported in Table 7, column 5.

- **“FirstStage_toyprice_foodprice.R”**: This program estimates the first stage equations and the reduced form equations using toy price and food price only as instrumental variables. The first stage results for material investments are reported in Table 5 (column 3).

- **"FirstStageFullyInteracted_toyprice_foodprice_conflict.R”**: This program estimates the investment functions where all the parameters (in addition to the intercept) are allowed to vary with the treatment and where toy price, food price and conflict are used as instruments. The results are reported in Appendix Table E.2.

- **“CobbDouglas_OLS_fullyinteracted.R”**: This program estimates the production function where all the parameters (in addition to the intercept) are allowed to vary with the treatment. The results are reported in Appendix Table E.3, column 2 (for socio-emotional skills).

- **“FirstStageForIVInteracted_toyprice_foodprice.R”**: This program estimates the first stage and reduced form equations that correspond to the production function where all parameters vary with treatment. It uses only toy prices and food prices as instrumental variables.

- **“CobbDouglas_IV_parsin_fullyinteracted.R”**: This program uses the first stage and reduced form estimates obtained from the previous program to estimate the production function for skills where all parameters vary with treatment and there is only one type of investment (material investment in the production function for cognitive skills and time investment in the production function for socio-emotional skills). The results for this specification of the production function for cognitive skills are reported in Appendix Table E.3, column 1.

- **"CobbDouglas_OLS_strata.R"**: This program estimates the production function for cognitive and socio-emotional skills by OLS where we also control for stratum fixed effects. These estimates are reported in Appendix Table E.7, columns 1 and 3.

- **"FirstStage_toyprice_foodprice_conflict_strata.R"**: This program estimates the first stage equations and the reduced form equations using toy price, food price, and conflict as instrumental variables and controlling for stratum fixed effects. The first stage results are reported in Appendix Table E.6.

- **"CobbDouglas_IV_strata.R"**: This program uses the first stage and reduced form estimates generated by the previous program to estimate the structural parameters of the production function controlling for stratum fixed effects. The results of this specification are reported in Appendix Table E.7, columns 2 and 4.

The programs above call a number of functions, which are defined in separate programs (one program per function) and named after the function they include. These are saved in the folder "code/R/functions".

e) Monte Carlo simulations

The paper reports results from a Monte Carlo aimed at assessing whether the estimates suffer from weak instrument bias. To replicate this Monte Carlo, one needs to perform the following steps:

- Estimate the covariance matrices A and B for material investment, time investment and cognitive skill using **"code/R/MC/ComputeCovarianceA.R"** and **"code/R/MC/ComputeCovarianceB.R"**. These covariances are reported in Appendix Table F.1.

- Generate 1000 simulated data samples using Covariance A and 1000 simulated data samples using Covariance B using **"code/R/MC/MASTER_MC.R"**. The user can choose option "A" or "B" at the top of this file to choose the relevant covariance and needs to choose option "generate = TRUE" to generate the data.

- Estimate the investment functions and production functions both by OLS and IV using the last step of **"code/R/MC/MASTER_MC.R"**, which calls the following relevant programs:

- **"CobbDouglas_OLS_MC.R"**: This program estimates the simpler version of the production function (with the treatment dummy, the child's cognitive skill at baseline, the mother's cognitive skill and the two investments) by OLS. The estimates of this model are reported in Table 8 and Appendix Table F.4.

- **"FirstStage_toyprice_foodprice_conflict_ MC.R"**: This program estimates the first stage equations and the reduced form equations using toy price, food price, and conflict as instrumental variables for the simpler model used in the Monte Carlo. The first stage results are reported in Table 8 and Appendix Table F.3. Standard errors for some of these coefficients are reported in Appendix Table F.2.

- **"CobbDouglas_IV.R"**: This program uses the first stage and reduced form estimates generated by the previous program to estimate the structural parameters of the simpler production function used in the Monte Carlo. The results of this specification are reported in Table 8 and Appendix Table F.4.

f) Exploratory factor analysis.

Results of the exploratory factor analysis reported in Appendix C are generated using the following programs:

- **"code/mplus/efa_XXX"** where XXX is "child0" for measures of child's skills at baseline; "child1" for measures of child's skills at follow up, "invest" for measures of investments at follow up; and "mother0" for measures of maternal skills at baseline: These programs are to be run in Mplus and perform an exploratory factor analysis of these groups of variables. These programs generate a scree plot and report the eigenvalues of the correlation matrix (used in the Kaiser rule). They also estimate the factor loadings on each of the measures, which we report in Appendix Tables C.2 through C.5

- **"code/stata/Table C.1"**: this Stata programs run the Velicer's MAP rule and Horn's parallel analysis to determine the number of factors underlying each group of measures.

g) Estimates of the model on other samples

The paper reports the estimates of our preferred specifications (i.e. Table 6, column 5 for cognitive skills and Table 7, column 1 for socio-emotional skills) on two additional samples:

- A sample excluding children who have received the nutritional supplementation component of the intervention, for which results are reported in Appendix Table E.4.

- A sample where tested effects have been removed from the raw measures entering the measurement system, for which results are reported in Appendix Table E.5.

To run the estimation for these two samples, one needs to perform the following steps:

- Generate the dataset of measures "measures.dta" and the corresponding bootstrapped datasets with the appropriate data using **"code/stata/prepare estimation sample.do"** and checking the desired option at the top of the file (pure = yes for the first sample; testerFE = yes for the second sample).

- Run **"code/R/MASTER.R"** with the desired options at the top of the file (puresample =1 for the first sample; testerFE = 1 for the second sample).

h) Estimates of the model allowing for treatment-specific estimates of the measurement system

The paper reports the estimates of our preferred specifications (i.e. Table 6, column 5 for cognitive skills and Table 7, column 1 for socio-emotional skills) based on a measurement system allowing for different loadings and intercepts between the treatment and control groups. To run this estimation, one needs to perform the following steps:

- Generate the dataset of measures "measures.dta" and the corresponding bootstrapped datasets with the appropriate data using **"code/stata/prepare estimation sample.do"** and checking the desired option at the top of the file (invar = yes). With this option, the measures are standardized to have mean 0 and standard deviation in the whole sample (instead of in the control group).

- Run **"code/R/MASTER.R"** with the desired option at the top of the file (invar =1). The estimates of the factor loadings in each group are reported in Appendix Tables E.8 and E.9. The estimates of the production functions using the corresponding data are reported in Appendix Table E.10.

i) A note on parallelising the estimation of the measurement system and factor scores

With a high number of bootstrapped samples, the estimation of the measurement system and factor scores can be time-consuming. An option to estimate the measurement model and the factor scores on subsets of the bootstrapped samples in parallel R sessions has been created so as to reduce the running time. To use this option, one needs to specify how many bootstrapped samples to run on each session (bsample), how many sessions there will be in total (pos), which session a particular window is running (node). Suppose we want to run the model on a total 1000 bootstrapped samples, breaking it down into 4 sessions each running the estimation on 250 samples: one would specific bsample=250, pos=seq(1,4,1), and node=1,2,3 or 4, depending which window is open. The window with node =1 will run the estimation on bootstrapped samples measures_b1 through measures_b250. The second one with node = 2 will run the estimation on samples measures_b251 through measures_b500. And so on. If one does not want to use this option, please specify bsample to be equal to the total number of bootstrap replications, pos=seq(1,1,1) and node=1.