**README – Replication Analysis of “Diversity in Schools: Immigrants and the Educational Performance of U.S.-Born Students,” Figlio et al. (2023), The Review of Economic Studies**

**Overview**

This document explains how to replicate the results (dataset construction, tables, and figures) of the paper “Diversity in Schools: Immigrants and the Educational Performance of U.S.-Born Students” by David Figlio, Paola Giuliano, Riccardo Marchingiglio, Paola Sapienza, and Umut Ozek. This document contains the following sections: 1) data availability and provenance statement; 2) computational requirements; 3) replication analysis; 4) list of tables and figures.

**1.** **Data availability and provenance statement**

The paper makes use of confidential administrative data from the Florida Department of Education and the Florida Department of Health. The data sharing agreement does not permit us to provide data directly to interested researchers. David Figlio is committed to helping researchers obtain access to the Florida administrative data, and over the past ten years he has assisted over one hundred researchers obtain access to confidential Florida data. To replicate the results researchers should obtain the data following the procedure outlined by the Florida Department of Education in their research data request portal (<http://www.fldoe.org/accountability/accountability-reporting/external-research-requests/>) and the Florida Department of Health in their research data request process (<http://www.floridahealth.gov/statistics-and-data/data-and-statistics/index.html> ).

The data from the Florida Departments of Education and Health are merged using name, date of birth, and social security number, so researchers interested in having the two agencies merge these datasets for them should be certain to describe the merge variables in their data applications. The Florida Department of Education was responsible for merging the datasets.

We certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

In addition to the confidential data, we also use data from Hofstede et al. (2010), Country codes (Agricultural Research Service, USDA), US Census Bureau Department of Commerce, FLDOE website.

The table below summarizes the data files used as inputs in our programs:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Source** | **Description** | **Notes** | **Provided** |
| FLDOE\_raw\_scores | FLDOE &Health department | Scores | Confidential | No |
| FLDOE\_raw\_demo | FLDOE &Health department | Demographic | Confidential | No |
| FLDOE\_raw\_mother | FLDOE &Health department | Maternal education | Confidential | No |
| FLDOE\_school\_points | FLDOE &Health department | Schools’  coordinates | Confidential | No |
| Crosswalk\_name\_id | FLDOE website | Crosswalk name-district id |  | Yes |
| Dist\_name\_coord\_2011 | US Census Bureau Department of Commerce | Shapefiles for 2011 |  | Yes |
| LTO\_Hofstede | Hofstede | LTO scores |  | Yes |
| ISO\_country\_codes | ARS, USDA | Countries’ classification |  | Yes |

The list of variables in each one of the raw data from the FLDOE and Health department are the following:

- **FLDOE\_raw\_scores**: k20\_edw\_id year institution\_id entry\_date withdrawal\_date lunch\_status grade rabsent student\_birth\_month student\_birth\_year country\_cd\_borned\_in prior\_kinder\_progr\_partp\_cd gender\_cd language\_have\_parents\_speaking racial\_ethnic\_cd schnum institution\_name district nincident suspended\_in suspended\_out suspended\_days\_out suspended\_days\_in incident suspended class\_id\_gr3 fcat\_grade\_level fcat\_read\_dev fcat\_math\_dev fcat\_read\_level fcat\_math\_level sfcat\_read sfcat\_math limited\_english\_proficiency\_cd primary\_excpt\_ind

- **FLDOE\_raw\_demo**: current\_academic\_year k20\_edw\_id test\_grade\_level subtest\_id dev\_scale\_score scale\_score achievement\_level achievement\_level\_2\_0 dev\_scale\_score\_2\_0

- **FLDOE\_raw\_mothers**: mother\_dob\_month mother\_dob\_year mother\_edcde older\_siblings gestation\_weeks k20\_edw\_id early\_prenatal\_care adequate married unique\_mom twinset mother\_birth\_country\_cde mother\_race mother\_hisp

**- FLDOE\_school\_points:** institution\_id district latcod loncod

**2. Computational Requirements**

**2.1 Software requirement**

Stata (code was last run with version 15)

The dependencies, installed locally by the code “macro\_dictionary.do”, are:

* spmap (as of 2018-01-18)
* groups (as of 2018-02-24)
* binscatter (as of 2013-11-24)
* outreg2 (as of 2014-08-17)
* ftools (as of 2019-11-18)
* reghdfe (as of 2019-11-18)
* unique (as of 2020-06-18)
* ivreghdfe (as of 2018-09-15)
* ivreg2 (as of 2022-05-10)
* ranktest (as of 2020-09-29)
* shp2dta (as of 2015-07-17)
* mif2dta (as of 2008-03-14)

**2.2 Memory and run-time requirement**

The code was run using a remote server, the Northwestern University Quest, using alternatively one of the two servers listed below, each with the characteristics outlined:

**Quest 10 - Interconnect: Infiniband EDR**

* Number of Nodes: 532 nodes with 27612 cores total, 52 cores per node
* Processor: Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz
* Memory: Per node (Per Core) 192 GB (3.7 GB), Type: DDR4 2666 MHz

**Quest 11 - Interconnect: Infiniband HDR compatible**

* Number of Nodes: 177 nodes with 11328 cores total, 64 cores per node
* Processor: Intel(R) Xeon(R) Gold 6338 CPU @ 2.0GHz
* Memory: Per node (Per Core) 256 GB (4 GB), Type: DDR4 2666 MHz

For further information: <https://www.it.northwestern.edu/departments/it-services-support/research/computing/quest/specs.html>

The approximate time required to run the code is 50 hours.

**3. Replication Analysis**

The replication package contains the following directories: (a) Data (b) Data Cleaning Codes (c) Analysis (d) Logs.

1. The Directory Data has the following subdirectories: (i) “/Data/1.Input.” where the confidential merged data obtained from FLDOE and Department of Health should be saved; (ii) “/Data/2.Build” where intermediate transformations data will be saved after running the codes; (iii) /Data/3.Output” where the dataset that will be used to produce the final tables and figures will be saved. The following table summarizes the full list of datasets.

|  |  |  |  |
| --- | --- | --- | --- |
| **Data file** | **Source** | **Notes** | **Provided** |
| /1.Input/FLDOE\_raw\_scores | FLDOE | Confidential | No |
| /1.Input/FLDOE\_raw\_math | FLDOE | Confidential | No |
| /1.Input/FLDOE\_raw\_mother | FLDOE | Confidential | No |
| /1.Input/FLDOE\_school\_points | FLDOE | Confidential | No |
| /1.Input/Shapefiles/Crosswalk\_name\_id | FLDOE site | - | Yes |
| /1.Input/Shapefiles/ Dist\_name\_coord\_2011 | US CBDC | - | Yes |
| /1.Input/LTO\_Hofstede | Hofstede | - | Yes |
| /1.Input/ISO\_country\_codes | ARS, USDA | - | Yes |
| /2.Build/country\_quality | - | Intermediate dta | No |
| /2.Build/DTA\_coordinates | - | Intermediate dta | No |
| /2.Build/ DTA\_LTO\_index\_to\_be\_merged | - | Intermediate dta | No |
| /2.Build/DTA\_LTO | - | Intermediate dta | No |
| /2.Build/DTA\_names | - | Intermediate dta | No |
| /2.Build/DTA\_school\_points | - | Intermediate dta | No |
| /2.Build/DTA\_T11\_intermediate | - |  |  |
| /2.Build/ DTA\_to\_be\_merged\_SESmgrnt\_exp | - | Intermediate dta | No |
| /2.Build/ DTA\_to\_be\_merged\_w\_LTOquartiles | - | Intermediate dta | No |
| /2.Build/FLDOE\_intermediate\_v1 | - | Intermediate dta | No |
| /2.Build/FLDOE\_intermediate\_v2 | - | Intermediate dta | No |
| /2.Build/Markov\_pred\_cum\_exp | - | Intermediate dta | No |
| /2.Build/Matrix\_exposures | - | Intermediate dta | No |
| /2.Build/pc\_mgrnt\_inst | - | Intermediate dta | No |
| /2.Build/pc\_mgrnt\_inst13 | - | Intermediate dta | No |
| /2.Build/quality | - | Intermediate dta | No |
| /2.Build/t4\_math\_short |  | Intermediate dta | No |
| /2.Build/t4\_math | - | Intermediate dta | No |
| /2.Build/t4\_read\_short | - | Intermediate dta | No |
| /2.Build/t4\_read | - | Intermediate dta | No |
| /2.Build/Temp\_trans\_mats/ | - | Folder with transition matrices | No |
| /3.Output/DTA\_tables\_paper | - | Output dta | No |
| /3.Output/DTA\_Tab7\_IV\_math | - | Output dta | No |
| /3.Output/DTA\_Tab7\_IV\_read | - | Output dta | No |
| /3.Output/iv\_sample\_for\_Fig6 | - | Output dta | No |
| /3.Output/T11/DTA\_T11\_means | - | Output dta | No |
| /3.Output/T11/ DTA\_T11\_numbobs\_Skeleton\_iyg\_`group’\_`subj’.dta | - | **(\*)** | No |

**(\*)** This last set of output datasets consists of 15 datasets used to produce Table 11. The groups (cfr. `group’ in the dta name) are: all, black, frl, nofrl, white. The subjects (`subj’) are: incident, math, read.

1. The directory “Data Cleaning Codes” contains:
   * “Master\_DO\_data\_transformation.do” (code that uses as input initial datasets and produces the datasets for the analysis, saved in the folder “/Data/3.Output” and the intermediate datasets, saved in the folder “/Data/2.Build”)
   * “macros\_dictionary.do” (containing a set of macros used by “Master\_DO\_data\_transformation.do” ).
2. The directory “Analysis” contains three subdirectories:
   * “/Analysis/DO files” contains the programs for reproducing the results: 1\_Main\_Paper\_tables.do and 2\_Main\_Paper\_figures.do.
   * “/Analysis/Tables” that contains the tables in excel.[[1]](#footnote-1)
   * “/Analysis/Figures” that contains the figures in pdf and a folder with temporary files for Figures production.
3. The directory “Logs” contains the log files for the do files, Master\_DO\_data\_transformation, 1\_Main\_Paper\_tables.do and 2\_Main\_Paper\_figures.do.

**4. List of Tables and Figures**

In the table are reported the output tables and figures and the corresponding DO files producing them.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Figure/**  **Table #** | **Program** | **Code**  **Line #** | **Output file** | **Notes** |
| ***TABLES*** | | | | |
| Table 1 | 1\_Main\_Paper\_Tables.do | 70 | 1\_Final\_analysis\_tables **(\*\*)** | .smcl file in Log folder |
| Table 2 | 1\_Main\_Paper\_Tables.do | 100 | 1\_Final\_analysis\_tables **(\*\*)** | .smcl file in Log folder |
| Table 3 | 1\_Main\_Paper\_Tables.do | 115 | 1\_Final\_analysis\_tables **(\*\*)** | .smcl file in Log folder |
| Table 4 | 1\_Main\_Paper\_Tables.do | 135 | Table4\_fcat\_math\_dev\_yg Table4\_fcat\_read\_dev\_yg | .xls files |
| Table 5A, 5B | 1\_Main\_Paper\_Tables.do | 155 | Table5A\_fcat\_math\_dev\_yg Table5A\_fcat\_read\_dev\_yg  Table5B\_fcat\_math\_dev\_yg Table5B\_fcat\_read\_dev\_yg | .xls files |
| Table 6A, 6B | 1\_Main\_Paper\_Tables.do | 180 | Table6A\_fcat\_math\_dev\_yg Table6A\_fcat\_read\_dev\_yg  Table6B\_fcat\_math\_dev\_yg Table6B\_fcat\_read\_dev\_yg | .xls files |
| Table 7 | 1\_Main\_Paper\_Tables.do | 210 | Table7\_fcat\_math\_dev\_yg  Table7\_fcat\_read\_dev\_yg  Table7\_1stSTAGE\_math  Table7\_1stSTAGE\_read | .xls files |
| Table 8 | 1\_Main\_Paper\_Tables.do | 295 | Table8\_col1\_math  Table8\_col2\_read  Table8\_col3\_math  Table8\_col4\_read | .xls files |
| Table 9 | 1\_Main\_Paper\_Tables.do | 325 | Table9\_col1\_fcat\_math\_dev\_yg\_all  Table9\_col1\_fcat\_read\_dev\_yg\_all  Table9\_col2\_fcat\_math\_dev\_yg\_nofrl  Table9\_col2\_fcat\_read\_dev\_yg\_nofrl  Table9\_col3\_fcat\_math\_dev\_yg\_frl  Table9\_col3\_fcat\_read\_dev\_yg\_frl  Table9\_col4\_fcat\_math\_dev\_yg\_white  Table9\_col4\_fcat\_read\_dev\_yg\_white  Table9\_col5\_fcat\_math\_dev\_yg\_black  Table9\_col5\_fcat\_read\_dev\_yg\_black | .xls files |
| Table 10 | 1\_Main\_Paper\_Tables.do | 425 | Table10\_col1\_fcat\_math\_dev\_yg\_all  Table10\_col1\_fcat\_read\_dev\_yg\_all  Table10\_col2\_fcat\_math\_dev\_yg\_nofrl  Table10\_col2\_fcat\_read\_dev\_yg\_nofrl  Table10\_col3\_fcat\_math\_dev\_yg\_frl  Table10\_col3\_fcat\_read\_dev\_yg\_frl  Table10\_col4\_fcat\_math\_dev\_yg\_white  Table10\_col4\_fcat\_read\_dev\_yg\_white  Table10\_col5\_fcat\_math\_dev\_yg\_black  Table10\_col5\_fcat\_read\_dev\_yg\_black | .xls files |
| Table 11 | 1\_Main\_Paper\_Tables.do | 485 | Table11\_number\_obs\_firstgen\_all\_incident  Table11\_number\_obs\_firstgen\_all\_math  Table11\_number\_obs\_firstgen\_all\_read  Table11\_number\_obs\_firstgen\_black\_incident  Table11\_number\_obs\_firstgen\_black\_math  Table11\_number\_obs\_firstgen\_black\_read  Table11\_number\_obs\_firstgen\_frl\_\_incident  Table11\_number\_obs\_firstgen\_frl\_math  Table11\_number\_obs\_firstgen\_frl\_read  Table11\_number\_obs\_firstgen\_nofrl\_incident  Table11\_number\_obs\_firstgen\_nofrl\_math  Table11\_number\_obs\_firstgen\_nofrl\_read  Table11\_number\_obs\_firstgen\_white\_incident  Table11\_number\_obs\_firstgen\_white\_math  Table11\_number\_obs\_firstgen\_white\_read  1\_Final\_analysis\_tables **(\*\*)** | For number of obs, refer to excel, .xls files; for means, refer to the .smcl file in Log folder |
| Table 12 | 1\_Main\_Paper\_Tables.do | 625 | Table12\_col1\_fcat\_math\_dev\_yg\_all  Table12\_col1\_fcat\_read\_dev\_yg\_all  Table12\_col2\_fcat\_math\_dev\_yg\_nofrl  Table12\_col2\_fcat\_read\_dev\_yg\_nofrl  Table12\_col3\_fcat\_math\_dev\_yg\_frl  Table12\_col3\_fcat\_read\_dev\_yg\_frl  Table12\_col4\_fcat\_math\_dev\_yg\_white  Table12\_col4\_fcat\_read\_dev\_yg\_white  Table12\_col5\_fcat\_math\_dev\_yg\_black  Table12\_col5\_fcat\_read\_dev\_yg\_black | .xls files |
| ***FIGURES*** | | | | |
| Figure 1 | |  |  |  | | --- | --- | --- | | 2\_Main\_Paper\_Figures.do | 12.5 KiB | 30-Jan-23 09:48:18 | | 65 | Figure1 | .pdf file |
| Figure 2 | |  |  |  | | --- | --- | --- | | 2\_Main\_Paper\_Figures.do | 12.5 KiB | 30-Jan-23 09:48:18 | | 140 | Figure2 | .pdf file |
| Figure 3A, 3B | |  |  |  | | --- | --- | --- | | 2\_Main\_Paper\_Figures.do | 12.5 KiB | 30-Jan-23 09:48:18 | | 160 | Figure3A  Figure3B | .pdf files |
| Figure 4 | |  |  |  | | --- | --- | --- | | 2\_Main\_Paper\_Figures.do | 12.5 KiB | 30-Jan-23 09:48:18 | | 200 | Figure4 | .pdf file |
| Figure 5 | |  |  |  | | --- | --- | --- | | 2\_Main\_Paper\_Figures.do | 12.5 KiB | 30-Jan-23 09:48:18 | | 220 | Figure5 | .pdf file |
| Figure 6A, 6B | |  |  |  | | --- | --- | --- | | 2\_Main\_Paper\_Figures.do | 12.5 KiB | 30-Jan-23 09:48:18 | | 240 | Figure6A  Figure6B | .pdf files |

**(\*\*)** In order to visualize the statistics for the tables 1, 2, 3, and 11 in the log file, it is just enough to look for the corresponding name of the table in the log file

**References**

Hofstede, Geert, Gert Jan Hofstede, and Michael Minkov. 2010. Cultures and organizations: software of the mind: intercultural cooperation and its importance for survival, 3rd ed., McGraw-Hill, New York.

Fusarium head blight country and keyword code, Agricultural Research Service, USDA: <https://www.ars.usda.gov/midwest-area/stpaul/cereal-disease-lab/docs/bibliographies/fusarium-head-blight-country-and-keyword-codes/>

Florida Department of Education, administrative data, 2002-2012:

<http://www.fldoe.org/accountability/accountability-reporting/external-research-requests/>

Florida Department of Health, administrative data, 1994-2002:

<https://www.floridahealth.gov/statistics-and-data/data-and-statistics/index.html>

Florida Department of Education, public data on Public Schools/Districts:

<http://web03.fldoe.org/schools/schoolmap_text.asp>

US Census Bureau Department of Commerce, TIGER/Line Shapefile, 2018, state, Florida, Current Unified School Districts Shapefile State-based:

<https://catalog.data.gov/dataset/tiger-line-shapefile-2018-state-florida-current-unified-school-districts-shapefile-state-based>

1. The tables are outputted in a raw excel version and they must be manually adjusted. The odd columns report the coefficient and the standard error in parenthesis, while the even ones report the same coefficient and the standardized beta coefficient in parenthesis below. Tables 1 to 3 should be retrieved by the log of the do files, since the statistics have been manually taken by the summary statistics without exporting them to excel. Tables 4 to 10 and Table 12 report all the columns of the raw output. Table 11 is created manually starting from the statistics saved in the log file. [↑](#footnote-ref-1)