Humans versus Chatbots: Scaling-up Behavioral Interventions to Reduce Teacher Shortages

Replication Files: ReadMe

Nicolás Ajzenman Gregory Elacqua Analía Jaimovich
Graciela Pérez-Núñez

Last update: 20 de septiembre de 2024

This package reproduces the figures and tables contained in the article "Humans versus Chatbots: Scaling-up Behavioral Interventions to Reduce Teacher Shortages". To replicate the tables in the paper, set the appropriate local directory where you saved the data files for each of the following STATA do files and R-script. Then run the code below to generate the necessary documents to reproduce the tables.

1. Overview

1.1. Software Requirements

- STATA: Used for all analyses, except for those of causal forest. Results were produced using v17.
- R: Only used for causal forest analysis.Results were produced using v4.3.2.

1.2. Replication Folder Structure

The replication folder contains the following sub-folders:

1. MainData: This folder includes the necessary datasets to replicate the figures and tables. It contains two main files:
    - `MainData.dta`: This is the primary database containing all the variables required for the analysis.
    - `CausalForestData.dta`: This is a reduced version of the main dataset used specifically for the Causal Forest Analysis, considering only human treated and control with interesed in education.

2. Dofiles : This folder contains the Stata Do files necessary to replicate every table in the main text. It contains 17 files:

   - `Main.do`: This is the main code that enables database cleaning and the creation of all the tables, which are referenced in 16 additional Do files.
   - `Table1.do - TableA9.do`: These are the 16 scripts that create each of the tables in the main and appendix document.

3. CausalForest: This folder contains the R files necesary to replicate Causal Forest analysis.

   - `Figure1.r`: Contains the R code to generate Figure 1 of the main document.
   - `Table7_IntermediateStepst.r`: Contains neccesary intermediate steps through the causal forest analysis, used for the creation of the Table 7.
   - `Figure2.r`: It contains the complete code to recreate Figure 2 of the main document.
   - `sample_cf_human.xlsx, education_rankfirst.xlsx, education_proportion.xlsx, education_listed.xlsx`: Intermediate results generated with the previos R script `Table7_IntermediateStepst.r`, for the final creation of Table 7.

4. Output: This folder contains all the final results of the tables and figures in the main document and its appendix.

Note: All result files ending with a letter (a, b, c) correspond to a panel of the respective table indicated in their name. They were subsequently combined into a single table in the text editor used for drafting the document.

2. Instructions for replicating results

All results can be reproduced by running the `files` listed in Section 1.2. In each of the files specifying `/Replication package"`, the directory where the downloaded replication package folder is located should be replaced. The replication process require a specific orde, running first the Causal Forest analyssis and then the Main Do File.

3. Preparation of the Database

To prepare the database presented in this replication package, a series of intermediate steps were carried out due to the integration of information from different sources in addition to the data generated by the experiment. All these intermediate steps were performed using a masked identifier for each student.

It's important to acknowledge that we're unable to share the original datasets or divulge the specific details of the data cleaning procedures, as they contain sensitive information and are not suitable for public distribution. However, the dataset provided (`MainData.dta`)

encompasses all the requisite variables for replication and analysis. Furthermore, the subsequent section elucidates the construction process of the primary variables utilized in the analyses.

3.1. Outcome variables creation

The following details how the outcome variables were created from the initial datasets received, which cannot be shared in this replication package. Firstly, it outlines how the selected careers were identified according to the grouping used throughout the article.

The initial dataset received from Department of Evaluation, Measurement and Educational Registration (DEMRE) provides the codes for the 10 applications made by students, arranged by the indicated preference. Using this, encoding was done following the CINE-UNESCO categories:

- **Education Outcomes** programs follow CINE-UNESCO[*] categories, including psychopedagogy and education programs.
- **Preschool Outcomes** programs is determined when the specialization of the degree is "Pedagogy in Preschool", one of the 11 specializations in the categorization of the DEMRE database.
- **Primary Outcomes** programs is determined when the specialization of the degree is "Pedagogy in Primary Education", one of the 11 specializations in the categorization of the DEMRE database.
- **Specialized Outcomes** programs is determined when the specialization of the degree is one of the 9 others specializations in the categorization of the DEMRE database.

To create the various types of outcomes, the following methodology was employed:

- **Application as first choice** is identified when the student indicated a career associated with the definitions previously mentioned as their first preference (out of the 10 possible).
- **Proportion of the choice set** is the quotient between the applications made in careers associated with the aforementioned definitions, relative to the total number of applications made (which could be less than 10).
- **Application at least once** corresponds to students who, within all their applications, present some reference to the previous categorization.
- **Enrolled** identifies students who have enrolled in a program within the aforementioned categories offered by a higher education institution whose admission process is centralized by the DEMRE.

[*]UNESCO (2012). International standard classification of education: Isced 2011, Comparative Social Research