README

Overview

Reproducibility package for: Group consulting continues to benefit firms after a decade: Experimental evidence from Colombian auto parts firms, *American Economic Review: Insights.*

The code in this replication package generates the 4 figures and 1 table in the paper, and 6 tables from the supplementary appendix. The analysis uses confidential data from four administrative data sources in Colombia. R code was used to construct some of the data files, and Stata to do the regressions and generate the figures. Log files are provided of all output given the difficulty of accessing these data in the datalab.

Data Availability and Provenance Statements

The data are not publicly available. This project uses confidential firm-level data from the Annual Manufacturing Survey (Encuesta Anual Manufacturera, EAM) for the 2010–2022 period (DANE, 2010-22), as well as the Technological Development and Innovation Survey (EDIT) (DANE, 2010/11-2019/2020). Access must be requested from the Departamento Administrativo Nacional de Estadística by emailing contacto@dane.gov.co and signing a confidentiality agreement. Once approved, access is granted only through the on-site processing room (Laboratorio de Microdatos) in Bogotá. All analysis is conducted within this secure environment using researcher-provided code (In this case, Stata codes). Outputs (such as log files and tables) cannot be exported directly and must be reviewed by DANE's technical team to ensure that no confidential firm-level information is disclosed. Only approved outputs can be released. The original source microdata are preserved indefinitely by DANE as part of its official statistical archive. They maintain the project folder with intermediate datasets, code and logs for one year. After that year, the raw survey files are still available, but the assembled folder would need to be rebuilt using our programs provided.

Our replication folder includes the Stata code and the approved outputs generated through this process. It also contains a Codebook that maps our variables to those in the DANE data. For confidentiality and disclosure review matters, the contact is spee@dane.gov.co (Sala de Procesamiento Especializado Externo).

In addition, we use two complementary data sources. First, export data from the Dirección de Impuestos y Aduanas Nacionales. These data come from exporters' declarations and include information on firm identifiers (NIT), export values, product-level detail, and destination markets. The website is: https://www.dian.gov.co/dian/cifras/Paginas/Bases-Estadisticas-de-Comercio-Exterior-Importaciones-y-Exportaciones.aspx. These data are publicly available on the website if you know a firm's tax identification number, and we used the confidential tax ids for our experimental sample to access the data for our firms. However, since

providing these data could enable firms to be identified, these data also cannot be disclosed in our replication package. These data will be preserved for a period of five years after publication following AEA policy.

Finally, we use administrative employment data from the Planilla Integrada de Liquidación de Aportes (PILA), managed by the Ministerio de Salud y Protección Social. Access to these data is not public and requires an institutional request. In our case, a formal request was submitted in cooperation with IPA, including the list of firms (by NIT), and the Ministry provided monthly information on formal employment for the firms included in the study. These data were used as a complementary source to track formal employment over time. These data will be preserved for a period of five years after publication following AEA policy.

In addition, baseline data (ManagementDatasetpublic.dta and EmploymentDatasetpublic.dta) from Iacovone et al. (2020) were merged with the PILA data.

As per AEA policy, the authors commit to preserving data and code for a period of no less than five years following publication of the manuscript, and to providing reasonable assistance to requests for clarification and replication

Statement about Rights

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

Summary of Availability

- ☐ All data **are** publicly available.
- ☐ Some data **cannot be made** publicly available.
- ☑ **No data can be made** publicly available.

Dataset list

Data file	Source	Notes	Provided
LTPila.dta	PILA	Confidential	No
nit_firmcode.dta	Authors	Confidential	No
ManagementDatasetpublic.dta	Iacovone et al. (2020)	Baseline	No
EmploymentDatasetpublic.dta	Iacovone et al. (2020)	Baseline	No
exit_panel_ext.dta	EAM	Confidential	No
eam_10_22.dta	EAM	Confidential	No
edit_12_20.dta	EDIT	Confidential	No
3.panel_ext_tech_year.dta	DIAN	Confidential	No

Computational requirements

Software Requirements

- Stata (code was last run with version 18).
- R 4.5.1 (2025-06-13 ucrt), platform x86_64-w64-mingw32/x64, for the two panel-construction programs. Package versions as installed on the laboratory workstation:

Package	**Version**
haven	2.5.5
tidyverse	2.0.0
readxl	1.4.5
sjlabelled	1.2.0
lmtest	0.9-40
rmarkdown	2.29

Run programs/00_setup_R_packages.R once to install them. It sets the repository to a dated CRAN snapshot (Posit Public Package Manager, 2025-11-17) and requests each package at the exact version listed above, then prints sessionInfo().

readxl is invoked as readxl::read_excel() in 0.EDIT_Panel rather than through library(), so an automated scanner will not detect it.

The two R programs are R Markdown notebooks; they can also be run with knitr::purl() followed by source().

Controlled Randomness

- ☑ No Pseudo random generator is used in the analysis described here.

Memory, Runtime, Storage Requirements

Summary

Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop machine:

- ☐ <10 minutes
- ☐ 10-60 minutes

- ☑ 1-2 hours
- ☐ 2-8 hours
- ☐ 8-24 hours
- ☐ 1-3 days
- ☐ 3-14 days
- ☐ > 14 days

Approximate storage space needed:

- ☐ < 25 MBytes
- ☐ 25 MB - 250 MB
- ☐ 250 MB - 2 GB
- ☐ 2 GB - 25 GB
- ☐ 25 GB - 250 GB
- ☐ > 250 GB
- ☑ Not feasible to run on a desktop machine, as described below.

Feasibility is not due to storage space constraints, but due to necessity of running analysis in Colombian data lab.

Hardware used for the original run

All programs were run on a Windows workstation inside DANE's SPEE. The figures below were read from the machine itself on 25 August 2026:

CPU: reported by the system as Intel64 Family 6 Model 85 Stepping 7, GenuineIntel (Intel Xeon, Skylake-SP / Cascade Lake generation), 4 logical processors

RAM: 24 GB (25,768,775,680 bytes)

Operating system: Windows 10 x64, build 19045

Storage: the assembled project folder occupied 1.54 GB across 458 files since each survey has different chapters (files) year by year with different extensions (.dta, .xlsx,…), comprising the extracted EAM and EDIT source files and the intermediate .dta datasets. Allow roughly 2 GB of free space to reproduce it.

No internet access; software and R packages are installed by DANE IT staff from an internal repository.

The analysis is not computationally demanding: it runs on a standard office workstation and requires no parallelisation, no cluster and no special hardware.

Approximate run time within the lab

Stata times are measured from the timestamps in the log files of the final run (17 November 2025):

Program	**Run time**
Table1_Ln.do	1 min 23 s
Table1_Levels_Survival.do	1 min 16 s
Table2.do	51 s
Table3_ExportEam.do	37 s
Table3_Innovation.do	54 s
Table3_Management.do	4 s
TableA4.do	36 s
TableA5.do	55 s
TableA6.do	35 s
Total, Stata	**≈ 7 minutes**
0.EAM_Panel_Construction.Rmd and 0.EDIT_Panel.Rmd (R)	**≈ 20 minutes**, dominated by reading the annual SAS files

Description of programs/code

There are three sets of R scripts:

00_setup_R_packages.R - installs the six packages the two R programs below need — haven, tidyverse, readxl, sjlabelled, lmtest and rmarkdown — from a dated CRAN snapshot, so the versions are pinned, and it prints sessionInfo() at the end.

0.EAM_Panel: This R script processes the raw data files from the Encuesta Anual Manufacturera (EAM) provided by DANE. The raw information is organized by year and by survey chapters (e.g., production, materials, employment, investment). The script performs the harmonization and merging steps required to build a consistent firm-level panel across years. This processing corresponds to the data transformation procedures performed in DANE's data laboratory. This data processing corresponds to the procedures carried out in DANE's data laboratory to transform the original EAM files into datasets. The resulting panel is then used to generate all tables and econometric results based on EAM information. The

two key output files are eam_10_22.dta and exit_panel_ext.dta, the panel used for firm perfomance variables, employment and firm exit analysis (with 1560 obs = 120 firms *13 years).The R script is converted to a text file so that it can be read on any computer.

0.EDIT_Panel: This R script processes the raw data from the Encuesta de Desarrollo e Innovación Tecnológica (EDIT) provided by DANE. The original survey data are delivered separately for each edition and divided into multiple chapters. The script carries out the harmonization and merging of all survey chapters across different EDITs, creating a unified firm-level panel. The key output file is edit_12_20.dta, containing the EDIT panel.

There are then different Stata do files used to estimate the treatment regressions, which are detailed below. Log files were used to record the output of these regressions given that all estimation had to take place in the data lab and the data themselves cannot be released. A file Figures2to4.do uses some of these estimates to construct figures in the main text.

Instructions to Replicators

The list of tables and programs below sets out the do files that are used to create each Table and Figure in the paper, and the Stata log file that contains the results that were obtained from running these files.

List of tables and programs

The provided code reproduces:

- ☐ All numbers provided in text in the paper
- ☑ All tables and figures in the paper
- ☐ Selected tables and figures in the paper, as explained and justified below.

Figure/Table no.	Program	Line no.	Output file	Log file	Notes
Figure 1	PILAAnalysis.do	121	Figure1.png		
Figure 2	Figures2to4.do	20	Figure2.png		
Figure 3	Figures2to4.do	218	Figure3.png		
Figure 4	Figures2to4.do	419	Figure4.png		
Table 1 (cols 1-3)	PILAAnalysis.do	292	Table1_1to3.csv	PILAAnalysis.log	
Table 1 (col 4)	Table1_Levels_Survival.do	215		logTable1_Levels_survival.log	
Table 1 (col 5)	Table3_Management.do	105		logTable3_Management_2018_2020.log	
Supplementary Appendix					
Table A1	Taken from Iacovone et al. (2023), Table 1				
Table D1 (col 1)	PILAAnalysis.do	325	TableD1PILA.csv	PILAAnalysis.log	creates estimates for Figure 2a.
Table D1 (col 2)	Table1_Levels_Survival.do	131		logTable1_Levels_survival.log	creates estimates for Figure 2b
Table D1 (col 3)	Table1_Ln.do	165		logTable1_Ln.log	
Table D1 (cols 4-7)	TableA4.do	135		logTableA4.log	creates estimates for Figure2c, 2d
Table D2	Table2.do	176		logTable2.log	creates estimates for Figure 3
Table D3	TableA5.do	254		logTableA5.log	
Table D4 (cols 1-2)	DIANcode.do	131	TableD4_1and2.csv	DIAN.log	creates estimates for Figure 4a & 4b
Table D4 (cols 3-4)	Table3_ExportEam.do	169		logTable3_ExportEam.log	
Table D4 (col 5)	Table3_innovation.do	129		logTable3_innovation.log	creates estimates for Figure 4c
Table D4 (col 6)	2.1Table2_LOG.do	332		logTable2_LOG.log	creates estimates for Figure 4d
Table E1	PILAAnalysis.do	355	TableE1.xls	PILAAnalysis.log	
Table E2	PILAAnalysis.do	420	TableE2.xls	PILAAnalysis.log	

References

Departamento Administrativo Nacional de Estadistica (DANE) Colombia - Encuesta Anual Manufacturera - EAM -2010-2022 https://microdatos.dane.gov.co/index.php/catalog/836

Departamento Administrativo Nacional de Estadistica (DANE) Colombia - Encuesta de Desarrollo e Innovación Tecnológica – EDIT – 2010/11-2019/20 - https://microdatos.dane.gov.co/index.php/catalog/651

Iacovone, Leonardo, William Maloney and David McKenzie (2020) "Replication package for: Improving Management with Individual and Group-Based Consulting: Results from a Randomized Experiment in Colombia", Zenodo doi: 10.5281/zenodo.4306042