

README — Reproducibility Package

When Digital Nudges Nag: Evidence on Program Take-up from a WhatsApp Intervention

María José Urbina • Andrés Moya • Sandra V. Rozo

Overview

This folder contains the reproducibility package for the Working Paper “When Digital Nudges Nag: Evidence on Program Take-up from a WhatsApp Intervention” by María José Urbina, Andrés Moya, and Sandra V. Rozo. If you run into any trouble running this code or reproducing results, please create an issue in this repository.

Data Availability Statement

1. Gran Encuesta Integrada de Hogares (GEIH)

Source	DANE (Departamento Administrativo Nacional de Estadística)
Year	2021
Filename(s)	area_diciembre.dta, cabecera_diciembre.dta, resto_diciembre.dta (and equivalent monthly files)
URL	https://microdatos.dane.gov.co/index.php/catalog/701
Access Date	March 2023
Access Type	Open — publicly available microdata through DANE’s microdata portal (ANDA)
License	Public use, anonymized microdata provided by DANE

Note: Microdata from the Gran Encuesta Integrada de Hogares (GEIH) 2021, collected continuously on a monthly basis. The dataset is organized in separate monthly files which are appended to construct a full-year dataset. For each month, files are organized into three geographic strata (*Área*, *Cabecera*, and *Resto*) and must be downloaded separately.

Merging strategy: Individual-level data are constructed by merging the following modules using household and individual identifiers (DIRECTORIO, SECUENCIA_P, ORDEN): Características generales (Personas), Desocupados, Fuerza de trabajo, Inactivos, Ocupados, Otras actividades y ayudas en la semana, and Otros ingresos. Household-level data (Vivienda y Hogares) are merged using household identifiers (DIRECTORIO, SECUENCIA_P) in a m:1 merge. The migration module (Migración) is merged at the individual level when available.

Sample: The analytical sample is restricted to Venezuelan-born individuals (`pais_nace == 3`) aged 18 and older. The construction code produces the file `GEIH.dta`, which is then combined with VenReps and ETPV data to construct `DataComparison.dta` (see Section 5 below).

Citation: DANE. 2021. “Gran Encuesta Integrada de Hogares” [dataset]. <https://microdatos.dane.gov.co/index.php/catalog/701>. Accessed March 2023.

2. VenReps Survey – Wave 1

Source	VenReps Project Team
--------	----------------------

Year	2021
Filename(s)	VenReps.dta
URL	https://www.venreps.com/
Access Date	March 2023
Access Type	Restricted — not publicly available
License	Restricted-use data. Access subject to approval by the data owners (VenReps Project Team).

Note: Proprietary microdata from the VenReps Survey, Wave 1, collected as part of the VenReps project, which studies Venezuelan migrants and host communities. The data contain individual- and household-level information, including demographic characteristics, migration history, labor market outcomes, and socioeconomic conditions.

Privacy restriction: Access to the raw VenReps microdata is restricted due to confidentiality agreements and data protection considerations. These data are **not included** in this replication package. Researchers interested in accessing the VenReps data should contact the VenReps Project Team directly at <https://www.venreps.com/>. The cleaning and harmonization code is included in the do-file for transparency.

Citation: *VenReps Project Team. 2021. “VenReps Survey – Wave 1” [dataset].*
<https://www.venreps.com/>. Accessed March 2023.

3. Registro Único de Migrantes Venezolanos (RUMV) – Municipal Aggregates

Source	Migración Colombia
Year	2022
Filename(s)	RUMV.dta
URL	Available upon request from Migración Colombia
Access Date	March 2022
Access Type	Restricted — shared directly by Migración Colombia for research purposes
License	Restricted administrative data provided by Migración Colombia

Note: Administrative data from the Registro Único de Migrantes Venezolanos (RUMV), provided by Migración Colombia. The dataset contains municipality-level aggregates on registered Venezuelan migrants. Municipal-level identifiers (e.g., municipality codes) are used to merge these data with other sources. A version of the dataset used for replication is included in the project replication package, subject to data protection considerations.

Citation: *Migración Colombia. 2022. “Registro Único de Migrantes Venezolanos (RUMV) – Municipal Aggregates” [dataset].* Accessed March 2022.

4. Project Survey Data – ETPV (December 2021)

Source	Innovations for Poverty Action (IPA)
Year	2021
Filename(s)	data_IPA.dta (raw), DataETPV.dta (clean)

URL	n.d.
Access Date	March 2022
Access Type	Restricted — included in replication package subject to data protection safeguards
License	Restricted-use data governed by data protection protocols and informed consent agreements

Note: Primary survey data collected in December 2021 as part of this research project, implemented by Innovations for Poverty Action (IPA). The Encuesta a Trabajadores y Población Venezolana (ETPV) is the study's own primary data collection. The dataset contains individual- and household-level information, including socioeconomic characteristics, program exposure, and outcomes of interest. The raw data (data_IPA.dta) are cleaned and processed using the do-file c_base.do to produce the analytical dataset DataETPV.dta.

Citation: *Innovations for Poverty Action (IPA). 2021. "Project Survey Data (December 2021)" [dataset]. Accessed March 2022.*

5. Constructed Dataset: DataComparison.dta

Filename(s)	DataComparison.dta
Construction script	01_build_DataComparison.do
Access Type	Partially restricted (see note)

Note: DataComparison.dta is a researcher-constructed dataset assembled from three sources: (1) the GEIH 2021, (2) the VenReps Survey Wave 1, and (3) the ETPV. It is used for Table 2 (Sample Comparability) in the paper. The construction code is provided in 01_build_DataComparison.do for full transparency.

The dataset retains only the variables that are comparably defined across all three surveys:

- **age:** Age in years (restricted to 18+)
- **sex:** Gender (1 = Male, 0 = Female)
- **years_education:** Years of completed education
- **income:** Log of monthly labour income (in millions COP), defined as $\log(\text{income}/1,000,000 + 1)$
- **health_reg__1:** Subsidized or special health regime (= 1)
- **health_reg__2:** Contributory health regime (= 1)
- **health_reg__3:** No affiliation or not reported (= 1)
- **unemployed:** Unemployment indicator (1 = Yes, 0 = No)
- **source:** Survey identifier ("GEIH", "PEP", "ETPV")

Privacy restriction: Because DataComparison.dta incorporates VenReps microdata, it is **not fully publicly available**. The GEIH and ETPV components are included in the replication package; however, the VenReps component cannot be redistributed without approval from the VenReps Project Team. The construction code (01_build_DataComparison.do) is fully provided so that researchers with access to all three sources can reconstruct the file.

Construction notes: (i) The GEIH sample is restricted to Venezuelan-born individuals (pais_nace == 3). (ii) From VenReps, only non-PEP-registered individuals are retained (pepramv_i == 0). (iii) The source variable is encoded alphabetically by Stata, so the numeric values of sample_1, sample_2, and sample_3 should be verified after running encode source, g(sample) by inspecting label list sample. (iv) All

monetary values are expressed in millions of Colombian pesos and log-transformed as $\log(\text{income} + 1)$ to handle zero values.

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.

I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permissions are documented in the LICENSE.txt file.

Instructions for Replicators

1. Download the repository folder. It includes the do-files and the datasets needed to run all tables and figures in the working paper.
2. Open **RepMaterials.do** and edit line 1 to reflect the path of the repository on your computer.
3. Execute **RepMaterials.do** to replicate all tables and figures. Note: datasets requiring restricted-access sources (VenReps) must be obtained separately before running the comparability table.

Computational Requirements and Runtime

Hardware	Windows 11 Home laptop, 16 GB RAM
Software	Stata (last run with version 18)
Runtime	Approximately 5 minutes
Storage	< 1 GB

Software Requirements

Stata version 18. All required packages are installed automatically by the master do-file RepMaterials.do at the start of the script.

Code Description

All analyses are implemented in Stata through two do-files:

- **c_base.do**: Assembles the final analytical dataset from the raw survey data (data_IPA.dta) and produces the cleaned dataset DataETPV.dta.
- **c_base_datacomparison** is the code to construct DataComparison.dta by harmonizing and appending the GEIH, VenReps, and ETPV sources
- **RepMaterials.do**: Master do-file. Sets the project paths, loads the required datasets, and reproduces all main tables, appendix tables, and figures reported in the paper. The code is organized sequentially by output type.

The master do-file installs required Stata packages, generates descriptive statistics and balance tables, treatment effect estimates, instrumental variables specifications, heterogeneous effects, attrition analysis, and robustness checks. Results are exported to the Outputs folder. Tables are adjusted for manuscript presentation using the excel2latex macro; figures are exported as publication-ready files.

List of Exhibits

- **Figure 1:** Constructed manually using ArcMap. The repository includes the shapefile for Colombian municipalities and departments.
- **Table 3:** Assembled from the outputs of Tables 3a, 3b, 3c, 3d, and 3e. The final version is generated in LaTeX using the excel2latex macro.
- **Figure A.3:** Constructed manually using ArcMap. The repository includes the shapefile and the RUMV data provided by Migración Colombia.
- **Table 4:** Assembled from the outputs Tab4_intend_rumv and Tab4_requested_ppt. Final table generated in LaTeX using the excel2latex macro.
- **Table A3:** Descriptive statistics table. Generated directly by RepMaterials.do and exported to the Outputs folder (previously mislabeled as A6 in earlier versions of the package; this has been corrected).
- **Table D1:** Constructed from outputs tabd1a and tabd1b, exported to LaTeX using the excel2latex macro.
- **Table E1:** Constructed from outputs e1 and e1b, exported to LaTeX using the excel2latex macro.

Folder Structure

DoFiles/	c_base.do: assembles DataETPV.dta from raw data and c_base_datacomparison constructs DataComparison.dta. RepMaterials.do: master script that generates all tables and figures.
Data/	DataETPV.dta: clean dataset for tables and figures. DataComparison.dta: comparability dataset (Table 2).
Raw_Data/	Data_IPA.dta: raw survey data from field work, DataComparison required data: ETPV, GEIH, VenReps.
Outputs/	Tables and figures included in the paper. Figures generated directly by Stata. Tables adjusted in Excel and exported to .tex using Excel2Latex.
LaTeX/	LaTeX manuscript and all files for compiling the final paper in PDF. All files produced by RepMaterials.do.
Shapefiles/	Shapefiles for Colombian municipalities and departments used in Figures 1 and A.3 (ArcMap).

Note on Reproducibility

All tables are reproducible provided that the same version of Stata is used. Stata's random number generator changed across versions (for example, between Stata 14 and 15), and set seed only guarantees the same sequence within the same version. If the code is always run using the same installation, the results will be identical.