README for Simone Bertoli, David McKenzie and Elie Murard (2025), "Migration, marriage and counterfactual spouses", accepted for publication on *Economic Development and Cultural Change*.

(version July 15, 2025)

Overview

The code in this replication package constructs the analysis files from four distinct data sources:

(i) IPUMS USA for the data for the 2017-2021 American Community Survey (ACS), for the 2008-2021 ACS (see Ruggles et al., 2025);
(ii) the Instituto Nacional de Estadística y Geografía (INEGI) for the 2020 Mexican Census;
(iii) the replication files from Mobarak et al. (2023);
(iv) the survey experiment conducted online by the Authors.

Data sources (i) to (iii) are not provided, as they are publicly available online and we describe below how they can be accessed, while the data source (iv) is provided and included in this replication package. Our replication package also contains all the (processed) data that are necessary to generate all the figures and tables that appear in the main text and in our online Appendix.

The source files have to be stored in the directory \1-Source. The do files in \3-Code\Data_Preparation generate the various datasets that are used in the analysis, which are stored in \2-Data. The do files in \3-Code\Data_Analysis generate the tables and figures included in the main text of the paper, and the tables and figures included in the online Appendix of our paper. The results of the analysis are stored in the directory \4-Results.

Replication requires storing all the files following the above instructions, adjust the file config.do, and then execute sequentially the two do files main_preparation.do and main_analysis.do. We have executed these three do files using Stata Version 18.5 on a computer whose operating system is Microsoft Windows.

This readme file is composed of the following three sections: Section 1, which describes all the data files that are used in the analysis, and Section 2, which described all the do files that execute the analysis. More precisely, Section 2.1 describes a single ancillary file whose role is just to specify the directory in which the files are stored; Section 2.2 describes the two do files that execute the all the do files that implement the preparation and the analysis of the data, and Sections 2.3 and 2.4 describe respectively the do files used in the data preparation and in the data analysis. Then, Section 3 lists (for each table and figure included in the main text or in the online Appendix) the do file and the data file(s) that are used to generate it.

Section 1. Data files

We detail here the five data files underlying our article:

ACS2017-2021.dta (not provided)

This data file includes the observations for the 2017-2021 American Community Survey (ACS). This data file can be accessed via the website https://usa.ipums.org/usa/. Access is subject to a mandatory but free registration that can be done via the following link: https://uma.pop.umn.edu/usa/user/new?return_url=https%3A%2F%2Fusa.ipums.org%2Fusa-

action%2Fmenu. The user should then login, and select the following sample: ACS 2021 5yr, and the following variables:

	variable name	storage type	variable label
•	year	int	Census year
•	multyear	int	Actual year of survey, multi-year ACS/PRCS
•	sample	long	IPUMS sample identifier
•	serial	double	Household serial number
•	cbserial	double	Original Census Bureau household serial number
•	numprec	byte	Number of person records following
•	subsamp	byte	Subsample number
•	cluster	double	Household cluster for variance estimation
•	strata	double	Household strata for variance estimation
•	pernum	int	Person number in sample unit
•	perwt	double	Person weight
•	sploc	byte	Spouse's location in household
•	related	int	Relationship to household head [detailed version]
•	sex	byte	Sex
•	age	int	Age
•	birthqtr	byte	Quarter of birth
•	marst	byte	Marital status
•	birthyr	int	Year of birth
•	marrno	byte	Times married
•	marrinyr	byte	Married within the past year
•	yrmarr	int	Year married
•	bpl	int	Birthplace [general version]
•	bpld	long	Birthplace [detailed version]
•	yrimmig	int	Year of immigration
•	yrsusa1	byte	Years in the United States
•	educ	byte	Educational attainment [general version]
•	educd	int	Educational attainment [detailed version

Then, the user should download: (a) the lines of code that are necessary to import the data in Stata format; (b) the data file. The data file at point (b) must be decompressed and saved. The do file at point (a) must be edited to include the location where the decompressed data file at point (a) has been saved. Once the data have been imported in Stata, then the file should be saved as ACS2017-2021.dta, and stored in the \1-Source folder.

ACS_2008_2021.dta (not provided)

This data file includes the observations for the 2008-2021 American Community Survey (ACS). This data file can be accessed via the website https://usa.ipums.org/usa/. Access is subject to a mandatory but free registration that can be done via the following link: https://uma.pop.umn.edu/usa/user/new?return_url=https%3A%2F%2Fusa.ipums.org%2Fusa-action%2Fmenu. The user should then login, and select the following samples: one-year sample from 2008 to 2021, and the following variables:

	variable name	storage type	variable label
•	year	int	Census year
•	sample	long	IPUMS sample identifier
•	serial	double	Household serial number
•	cbserial	double	Original Census Bureau household serial number
•	hhwt	double	Household weight
•	expwth	double	Experimental housing unit weight

- cluster double Household cluster for variance estimation
- strata double Household strata for variance estimation
- gq byte Group quarters status
- pernum int Person number in sample unit
- cbpernum int Original Census Bureau person number in sample unit
- perwt double Person weight
- expwtp double Experimental person weight
- sex byte Sex
- age int Age
- marst byte Marital status
- birthyr int Year of birth
- marrno byte Times married
- marrinyr byte Married within the past year
- yrmarr int Year married
- bpl int Birthplace [general version]
- bpld long Birthplace [detailed version]
- yrimmig int Year of immigration
- yrsusa1 byte Years in the United States
- yrsusa2 byte Years in the United States, intervalled
- migrate1 byte Migration status, 1 year [general version]
- migrate1d byte Migration status, 1 year [detailed version]
- migcounty1 int County of residence 1 year ago

Then, the user should download: (a) the lines of code that are necessary to import the data in Stata format; (b) the data file. The data file at point (b) must be decompressed and saved. The do file at point (a) must be edited to include the location where the decompressed data file at point (a) has been saved. Once the data have been imported in Stata, then the file should be saved as ACS_2008_2021.dta, and stored in the \1-Source folder.

Personas00.csv (not provided)

This data file (in .csv format) includes the individual observations from the long form of the questionnaire of the 2020 census of the Mexican Population that was administered by the INEGI to a subset of 20 percent of households. Access is free and not subject to a registration. The data can be accessed here: https://www.inegi.org.mx/programas/ccpv/2020/#microdatos, and the user should follow the path "Archivos para descarga/Muestra (cuestionario ampliado)/Bases de datos/", and then clicking on the line CSV corresponding to "Estados Unidos Mexicanos". This will result in the download of the compressed folder Censo2020_CA_eum_csv.zip (464 Mb). Then, the used should decompress the file Personas00.csv included in the folder, and store it in the \1-Source folder.

applicant_final.dta (not provided)

This data file belongs to the replication package of Mobarak et al. (2023). The replication package can be accessed here: https://www.openicpsr.org/openicpsr/project/174281/version/V1/view Access is subject to a mandatory but free registration that can be done via the following link: https://login.icpsr.umich.edu/realms/icpsr/protocol/openid-connect/auth?client_id=openicpsr-web-prod&response_type=code&login=true&redirect_uri=https://www.openicpsr.org/openicpsr/oauth/callback. The user should then login, and download, from the data folder of the replication package, the file applicant_final.dta (1.8 Mb), which should then be stored in the folder the \1-Source.

online_survey_dataset.dta (provided)

This data file contains the data of the online survey experiment conducted by the Authors in 2025 on a sample of 1,007 unmarried men aged 18 to 25 that were residing in Mexico.

Section 2. Do files

We detail here all the do files used in the data preparation and in the data analysis, which are all included in the replication package.

Section 2.1 Ancillary file

config.do

This do file allows the user to specify the main directory of the files, and it installs (if necessary) four packages in Stata.

Section 2.2 Main files

main_preparation.do

This do file executes, in approximately 30 minutes, all the do files of the data preparation. This do file was executed for the last time on June 14, 2025.

main_analysis.do

This do file executes, in approximately 3 minutes, all the do files of the data analysis. This do file was executed for the last time on July 21, 2025.

Section 2.3 Data preparation

The following do files are used for data preparation:

build_ACS2017_2021.do

This do file takes as input the dataset **ACS2017-2021.dta**, and it creates the variables necessary for the analysis, and it then saves the new data file **ACS2017-2021_finalready.dta** as output in the folder \2-Data.

Import_mexcen2020.do

This do file takes as input the data file **Personas00.csv**, it imports it in Stata, and it creates the variables necessary for the analysis, and it then saves the new data file **Mexcensus2020_final.dta** as output in the folder \2-Data.

build_combined_ACS_Mexcensus.do

This do file takes as input the dataset **Mexcensus2020_final.dta** and **ACS2017-2021_finalready.dta** combines and harmonizes the two, and it then saves a new data file **ACS_Mexcensus_combined.dta** as output in the folder \2-Data.

preparation_mobaraketal.do

This do file takes as input the dataset **applicant_final.dta**, and it creates the additional variables necessary for the analysis, and it then saves the new data file **Mobarak_replication_data.dta** as output in the folder \2-Data.

build_ACSpanel_0821.do

This do file takes as input the data file **ACS_2008_2021.dta** to select the observations related to immigrants born in developing countries in the ACS 2008-2021. It then saves the new data file **ACS_2008_2021_coh0712.dta** as output in the folder \2-Data.

varDef.do

This do file is based on a file (with the same name) included in the replication package (programs folder) of Mobarak et al. (2023). This do file includes the definition of the variables that are necessary for the analysis of the data.

Section 2.4 Data analysis

The following do files are used for the data analysis in the main text and in the online Appendix (the name of each file refers to the corresponding table(s) or figure(s) that it generates):

Table 1.do

This do file uses as input the dataset **ACS2017-2021_finalready.dta**, and it creates the file **Table1.smcl** in the directory \4-Results.

Table 2.do

This do file uses as input the dataset **ACS2017-2021_finalready.dta**, and it creates the file **Table2.smcl** in the directory \4-Results.

Table 3 & Table 4.do

This do file uses as input the dataset **ACS_Mexcensus_combined.dta**, and it creates the files **Table3.smcl** and **Table4.smcl** in the directory \4-Results.

Table 5.do

This do file uses as input the dataset **online_survey_dataset.dta**, and it creates the file **Table5.smcl** in the directory \4-Results.

Table 6 & Table C7 & Table C8.do

This do file uses as input the dataset **online_survey_dataset.dta**, and it creates the files **Table6.smcl**, **TableC7.smcl** and **TableC8.smcl** in the directory \4-Results.

Table 7.do

This do file uses as input the dataset **online_survey_dataset.dta**, and it creates the file **Table7.smcl** in the directory \4-Results.

Table C.6.do

This do file uses as input the dataset **online_survey_dataset.dta**, and it creates the file **TableC6.smcl** in the directory \4-Results.

Table A1-A.5.do

This do file uses as input the dataset **ACS2017-2021_finalready.dta**, and it creates the tables **TableA1.smcl** to **TableA5.smcl** in the directory \4-Results.

Figure1_figureB.1.do

This do file uses as input the dataset **ACS2017-2021_finalready.dta,** and it creates the files **Fig1_panelA.eps**, **Fig1_panelB.eps**, **FigB1_panelA.eps** and **FigB1_panelB.eps** in the directory \4-Results.

Figure2_figureB.2.do

This do file uses as input the dataset **ACS2017-2021_finalready.dta**, and it creates the files **Fig2_panelA.eps** and **Fig2_panelB.eps**, and **FigB2.eps** in the directory \4-Results.

FigureB.3_FigureB.4.do uses as input the dataset **ACS_collapsed_2008_2021.dta,** and it creates **FigB3_panelA.eps**, **FigB3_panelB.eps**, **FigB4_panelA.eps** and **FigB4_panelB.eps** in the directory \4-Results.

Section 3 Production of each table and figure in the main text and in the online Appendix

	Do-file	Stored result
Main text		
Table 1. Immigrants from developing countries	Table 1.do	Table1.smcl
Table 2. Immigrants from selected developing countries	Table 2.do	Table2.smcl
Table 3. Migration and spouse characteristics	Table 3 & Table 4.do	Table3.smcl
Table 4. Variance of spouse characteristics (residuals)	Table 3 & Table 4.do	Table4.smcl
Table 5. Expectations about marriage and future spouse in two scenarios: migrating versus staying	Table 5.do	Table5.smcl
Table 6. Effects of priming on migration intentions	Table 6 & Table C.7 & Table C.8.do	Table6.smcl
Table 7. Impact on marriage and spouse characteristics using Mobarak et al. (2023)'s data (IV estimates)	Table 7.do	Table7.smcl
Figure 1. Change in marital status with years of migration	Figure1_figureB.1.do	Fig1_panelA.eps Fig1_panelB.eps
Figure 2. Migrants married before migration: co-residence with spouse and separation	Figure2_figureB.2.do	Fig2_panelA.eps Fig2_panelB.eps
Online Appendix		
Table A.1 Immigrants from Mexico	Table A1-A.5.do	TableA1.smcl
Table A.2 Immigrants from India	Table A1-A.5.do	TableA2.smcl
Table A.3 Immigrants from China	Table A1-A.5.do	TableA3.smcl
Table A.4 Immigrants from the Philippines	Table A1-A.5.do	TableA4.smcl
Table A.5 Immigrants from the MENA region	Table A1-A.5.do	TableA5.smcl
Figure B.1 Robustness: Change in marital status with years of migration	Figure1_figureB.1.do	FigB1_panelA FigB1_panelB
Figure B.2 Mexican male migrants married before migration	Figure2_figureB.2.do	Fig2_panelB.eps
Figure B.3 Return migration with years since migration	FigureB.3_FigureB.4.do	FigB3_panelA.eps FigB3_panelB.eps
Figure B.4 Change in marital status adjusted for return migration	FigureB.3_FigureB.4.do	FigB4_panelA.eps FigB4_panelB.eps
Table C.6 Comparison of survey sample with Mexican Population Census among 18-25 unmarried men	Table C.6.do	TableC6.smcl
Table C.7 Effect of priming on migration intentions	Table 6 & Table C.7 & Table C.8.do	TableC7.smcl
Table C.8 Correlations between expectations of future spouse and migration intentions	Table 6 & Table C.7 & Table C.8.do	TableC8.smcl

References

Mobarak, Ahmed Mushfiq, Iffath Sharif, and Maheshwor Shrestha. 2023. "Returns to International Migration: Evidence from a Bangladesh-Malaysia Visa Lottery." *American Economic Journal: Applied Economics* 15 (4): 353–88.

Ruggles, Steven Sarah Flood, Matthew Sobek, Daniel Backman, Grace Cooper, Julia A. Rivera Drew, Stephanie Richards, Renae Rodgers, Jonathan Schroeder, and Kari C.W. Williams. 2025? IPUMS USA: Version 16.0 [dataset]. Minneapolis, MN: IPUMS, 2025. https://doi.org/10.18128/D010.V16.0