Replication of a Research Claim from Liang et al. (2018),
from *Journal of Political Economy*

Replication Team: Andrew Tyner, Nathan Fiala, and Victor Volkman

Research Scientist: Anna Abatayo

Action Editor: Kim Peters

Independent Reviewers
(add name below when you initiate review, comment “DONE” on your name when you finish):

Reviewer #1: Daniel Mallinson DONE

Reviewer #2: [Christine Y.M. Fong]DONE

Reviewer #3: William Chopik [Done]

Review Period: August 17 - August 24

View-only links to: Original Paper, Original Materials, Replication Data, Replication Analysis

Privacy Statement: Other teams are making predictions about the outcomes of many different studies, not knowing which studies have been selected for replication. As a consequence, the success of this project requires full confidentiality of this peer review process. This includes privacy about which studies have been selected for replication and all aspects of the discussion about these replication designs.

Instructions for Data Analysts

The preregistration for this replication study was started by a separate team of researchers who were responsible for identifying data sources and constructing them into a replication dataset(s) for your use in the analysis. They have completed sections 1-13 of the preregistration below, and included additional materials in the OSF project that document how the dataset was constructed.

In cases where all of the underlying data sources were able to be freely shared and posted, the constructed dataset(s) have been posted to the OSF as well, which you are free to use in designing the analysis plan (see below for details). In cases where some or all of the data sources could *not* be freely shared or posted, the replication dataset(s) are not provided on the OSF. Rather, you will need to follow the instructions and code to first reconstruct the datasets, and then proceed with your work. In such cases, the team responsible for creating the dataset(s) has provided summary statistics in the OSF that correspond to the constructed datasets, so you can verify that the datasets you create match what they intended.

You'll be responsible for filling out sections 16-25 of the preregistration below. Before you do so, **please review the original study, sections 1-15 of the preregistration, and the materials provided on the OSF**, so that you are familiar with all of the decisions that have been made to date. In many cases, the 'data preparer' will have left you instructions and suggestions on how the provided data can be used in the analysis, as well as idiosyncrasies and discrepancies in the data that you should be aware of. The data preparers have tried to be thorough in including all variables that you might need, but please keep in mind the following:

- Some of the variables included in the constructed dataset(s) may not be needed in the final analysis, so please do not feel the need to necessarily use all of the provided variables.
- Some of the variables needed might have mistakenly been excluded from the constructed datasets. If you find that this is the case, please let Andrew or Anna know, and they will work with you to supplement the datasets as needed.

For these secondary data replications, we would like the analysis plan to be completed before the preregistration goes through review, so that after review, the only remaining steps are registration and running the analysis code on the full datasets. To facilitate that, we are asking that you include in section 19 a link to the code you will use that takes the constructed dataset(s) provided to you and produces the focal analysis (including all of the cleaning, merging, and transforming required). When developing your analysis plan and code, please randomly sample 5% of the data for use in your work, and **do not use the rest of the data until it is time to run the final analysis**. In section 19, you will find a statement that we are asking you to bold that confirms you've only used 5% of the data when developing and testing your code. If this approach will not work for any reason, please let Andrew or Anna know and disclose deviations from this plan somewhere in the preregistration.

- In cases where we are providing you a complete dataset, you can just sample out 5% of the observations and hold the rest out until you are ready to perform the final analysis.
- In cases where we are providing you multiple datasets that need to be combined prior to analysis, please sample out 5% of the observations in whatever way is most sensible.
    - For example, in cases where each dataset contains complete observations on its own (a typical 'row bind' situation), it makes the most sense to sample out 5% of each dataset separately and then combine them together to develop and test your code.
    - In cases where datasets need to be merged in order to create complete observations (a typical 'column bind' situation), it makes the most sense to merge the separate datasets

into a full dataset first, and then sample out the 5% before proceeding with the rest of the analysis code.

- We leave the decision on how to sample out the random subset of data to you, so long as (a) you are not performing any analyses on the complete dataset until after your study is registered and (b) whatever decision you make is documented in the preregistration.

Finally, in cases where the replication data combines observations from the original study with observations that were not used in the original study (what we are calling 'hybrid replications'), please perform two analyses (details immediately below). This will likely require you to subset your data into the two groups described immediately below, based on the description of the original analysis provided in the study.

- When the 'new' data alone can clear the minimum power threshold, please perform one analysis that combines all available data, and a second that only uses the 'new' data. Please make sure both analyses are documented (with code) in section 19 below.
- When the 'new' data alone *cannot* clear the minimum power threshold, please perform one analysis that combines all available data, and a second that only uses the old data. Please make sure both analyses are documented (with code) in section 19 below.

Please contact Andrew or Anna if you have any questions. After you've completed the remaining sections of the preregistration and uploaded all the necessary materials to the OSF, please contact the SCORE coordinators regarding next steps.

Preregistration of Liang_JournPoliEco_2018_q8xv

Existing Data Replication

Study Information

1. Title (provided by SCORE)

RR TEAM INSTRUCTIONS: *This has been determined by SCORE.*

Replication of a research claim from Liang et al. (2018) in *Journal of Political Economy*.

2. Authors and affiliations

RR TEAM INSTRUCTIONS: *Fill in the names and affiliations of your team below.*

Nathan Fiala [Leads replication and analysis][1]
Andrew Tyner [Data identification and preparation][2]
Victor Volkman [Student researcher assisting with analysis and troubleshooting code][3]

1 Department of Agricultural and Resource Economics,University of Connecticut, Storrs, CT
2 Center for Open Science, Charlottesville, VA
3 Department of Economics,University of Connecticut, Storrs, CT

3. Description of study (provided by SCORE)

RR TEAM INSTRUCTIONS: *This description has been provided by SCORE. Please review and make a SCORE project coordinator aware of any edits, additions, and corrections you would suggest to the paragraph. You are free to add additional descriptions of your project in a separate paragraph.*

The claim selected for replication from Liang et al. (2018) is that the number of entrepreneurs as a fraction of the workforce decreases with the country's median age. This reflects the following statement from the paper's abstract: "A one standard deviation decrease in a country's median age increases new business formation by 2.5 percentage points, which is about 40 percent of the mean rate." [Tested with] country-Year-Level entrepreneurship rate regression. Table 2 reports reduced-form regressions at the country level of aggregation. Portion of Table 2 selected is Column 3. Focal independent variable is "Median age (ages 20-64)". [Coefficient of "Median age (ages 20-64)" is -0.007 with standard errors clustered at the country level of -0.001, significant at 1 percent.] Using the estimates from column 3 of table 2, a one standard deviation decrease in median age (equal to 3.5 years in 2010) results in a 2.5 percentage point

increase in the entrepreneurship rate, which is over 40 percent of the mean entrepreneurship rate across countries (equal to 0.061 in 2010).

4. Hypotheses (provided by SCORE with possible Data Analyst additions)

RR TEAM INSTRUCTIONS: *The focal test for SCORE is indicated as H*. If you will test additional hypotheses (or use alternate analyses) that help you to evaluate the claim your replication/reproduction is testing, number them H1, H2, H3 etc. (You can place H* in the list wherever makes sense). Please make sure that any additional hypotheses are logical deductions/operationalizations of the selected SCORE claim or are necessary to properly interpret the focal H* hypothesis. Research that is outside this scope should be described in a separate preregistration.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Are the listed hypotheses specific, concise, clearly testable, and specified at the level of operationalized variables?*
- *Are hypotheses identified as directional or non-directional, and, if applicable, have the direction of hypotheses been stated? (Example: "Customers' mean choice satisfaction will be higher in the CvSS architecture condition than in the standard attribute-by-attribute architecture condition.")*
- *Does the list of hypotheses/tests indicate whether additional hypotheses are taken from the original study or modified/added by the team?*

H*: The entrepreneurship rate in a country is negatively associated with the country's median age.

Design Plan

5. Study type

NOTE: *The study type selected should be based on the data collected for the replication, and not necessarily the data used in the original study.*

- Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.
- **Observational Study - Data is collected from study subjects that are not randomly assigned to a treatment. This includes surveys, natural experiments, and regression discontinuity designs.**
- Meta-Analysis - A systematic review of published studies.
- Other

6. Blinding

RR TEAM INSTRUCTIONS: *Select any/all of the below that apply for your study by bolding them. You will give a longer description in the next question.*

- **No blinding is involved in this study.**
- For studies that involve human subjects, they will not know the treatment group to which they have been assigned.
- Personnel who interact directly with the study subjects (either human or non-human subjects) will not be aware of the assigned treatments. (Commonly known as “double blind”)
- Personnel who analyze the data collected from the study are not aware of the treatment applied to any given group.

[QUESTION 6 - BOLD YOUR RESPONSE ABOVE]

7. Blinding

RR TEAM INSTRUCTIONS: *Since all existing data replications are based on data that has already been collected, in most cases it will not be necessary to comment on participant blinding. In the rare instance when an existing experiment is being re-analyzed for an existing data replication and blinding is a relevant consideration, please provide below any details regarding blinding that are important for a reviewer to be aware of.*

No blinding was involved to the secondary data collectors’ knowledge.

8. Study Design

RR TEAM INSTRUCTIONS: *Please describe how data was collected in the original study and how it compares to the data that was selected for the replication attempt. Explain why the data selected for the replication study is suitable for a replication and if any substantial deviations exist between the two.*

If the data used in the replication combines observations from the original study with new observations (e.g. if the data selected for the replication attempt comes from the same longitudinal survey as the original study), describe how 'original' and 'new' observations relate to each other and an estimate for what proportion of the final dataset's observations will be comprised of original vs. new observations.

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify the unit of analysis?*
- *Does the preregistration provide sufficient detail about how the data selected for the replication attempt deviates from or is congruent with the data employed in the original study?*
- *Does the preregistration describe whether and how 'original' and 'new observations' are combined together for the replication dataset?*

The focal analysis in the original study relies on data from two sources: the Global Entrepreneurship Monitor (GEM) and the U.S. Census Bureau's International Database, with a sample period of 2001-2010.

The replication study will rely on the same data sources, but extend forward in time to end with the latest year of available GEM data (2001-2016). To fully assess the replicability of the conclusions of the original study, regressions will be run seperately on the entire 16 year period, only the data from the original study, and only the data not included in the original study. All of the measures needed to reproduce the variables used in the focal analysis are available in each year of the GEM and Census data.

As indicated in section 13 below, approximately half of the observations (country-years) included in the replication dataset are from the original analysis (2001-2010) and half are from years not included in the original analysis (2011-2016).

9. Randomization (free response)

RR TEAM INSTRUCTIONS: *If the variables used for this replication attempt were randomized, state how they were randomized, and at what level.*

The independent variable is a country-level variable, so randomization does not apply.

Sampling Plan

This section describes how the data sources for the replication were selected, how they were prepared into a replication dataset, and the number of observations that will be analyzed from these data. Please keep in mind that the data described in this section are the actual data used for analysis, so if you are using a subset of a larger dataset, please describe the subset that will actually be used in your study.

10. Existing data (multiple choice question, provided by SCORE)

1.1.1. Registration prior to creation of data
1.1.2. Registration prior to any human observation of the data
1.1.3. Registration prior to accessing the data
1.1.4. Registration prior to analysis of the data
1.1.5. Registration following analysis of the data

11. Explanation of existing data

NOTE: *For replications that rely on existing data sources, this question refers to the data that will be used for the replication analysis (i.e. the final replication dataset), and not (a) the data from the original study or (b) the data sources accessed to construct the replication dataset. Since no new data will be created for 'existing data replications,' 1.1.1 should never be selected. Since all analyses will occur after registration, 1.1.5 should also never be selected.*

The datasets referenced above -- GEM datasets for 2001-2016, and Census Bureau data for the same years -- have been accessed, cleaned, and merged prior to registration. Variables were selected based on their expected relevance to the replication analysis. None of the variables were selected because of their likelihood (or not) of leading to a confirmatory result. A full hybrid analysis was performed on the dataset as a whole prior to registration.

12. Data collection procedures

RR TEAM INSTRUCTIONS: *Please describe the process for constructing the replication dataset in as much detail as you can. The sections below should be used to provide the following information:*

- *Which variables are needed from the original study to perform a good-faith, high-quality replication.*
- *Which data sources were used, why they were selected, any deviations between the original study design and the replication study design that these selections present, and the procedures used to access the data.*

- *Which of the variables from the original study are available in the replication data sources, including relevant details about each measure.*
- *The procedure for creating the replication dataset, in both narrative and script form.*
- *A data dictionary that documents each variable included in the replication dataset.*

In the sections below, please provide links to the original materials whenever possible -- including descriptions of the original datasets and corresponding codebooks. If materials can be shared on the OSF, please do so, and provide view-only links to those materials.

Specific points to keep in mind for reviewers:
- *Does the preregistration describe which data sources were selected for the replication study and why each is suitable?*
- *Does the preregistration make clear how the data sources were used to construct the replication dataset?*

(a) Data Needed

RR TEAM INSTRUCTIONS: *List below the datasets and variables the original author used to analyze the focal claim. Include details regarding the sample size, waves or years used, and other details pertinent to finding an existing dataset for replication. Please include page numbers when excerpting from the original article. If possible, categorize the list of variables as one of the following: dependent variable, focal independent variable, control variable, or sample parameters/clustering variable. Finally, include the sample size of the original study's focal analysis, if it is available.*

Dependent Variable(s)

Entrepreneurship rate
- Global Entrepreneurship Monitor [GEM], Adult Population Survey [2001-2010]
- "There are a number of different entrepreneurship rates that are reported in the GEM. For most of the empirical analysis, entrepreneurship is defined as "manages and owns a business that is up to 42 months old and pays wages."" (p. S158)
- Though not made explicit in the paper, the focal entrepreneurship rate variable is constructed as the proportion of respondents in each country-wave who '[manage] and owns a business that is up to 42 months old' *among those aged 20-64*.
    - The 'pay wages' criterion appears to be a feature of the variable used to measure this concept in the original study, rather than a separate variable.
- Additional notes: "The survey was carried out in 82 countries in various years from 2001 to 2010. Many countries were surveyed multiple times, but the panel is an unbalanced one. (Refer to table B1 in app. B for detailed information on countries, timing of the survey, and sample size in the GEM data.)" (p. S159)

Focal Independent Variable(s)

Median Age

- "The population statistics come from the US Census Bureau's International Data Base (IDB). The IDB is updated routinely and contains estimates and projections for over 200 countries and areas of the world. The IDB provides population counts by age from age 0 to 100 plus for each of the countries every year." (p. S160)
    - "Columns 3 and 4 repeat the analysis [regress the country's entrepreneurship rate] but replace r with the median age in the country. [fn 30: The median age is calculated only for those 20–64 years old.]" (p. S165) Note: *r* is the shrinkage parameter relating the size of one age cohort to another.

Additional variables

Weights

- "Observations are weighted by the number of individuals who make up each country-year cell." (p. S166 [Table 2 footnote])

Years

- "Year dummies are included in all regressions" (p. S166 [Table 2 footnote])

Country

- "Standard errors clustered at the country level are in brackets." (p. S166 [Table 2 footnote])

The focal analysis [Column 3 of Table 2] relies on 393 observations.

(b) Data Access

RR TEAM INSTRUCTIONS: *Describe below the data sources that will provide the replication variables. Include information such as the name of the data source (e.g., Indonesian Family Life Survey), the description and link of the data source, and the waves needed to create a final replication dataset.*

Also describe the process for accessing the data sources that will be used to create the final replication dataset; specify how long long it took for the registration to be approved and what information was required (e.g., writeup of the purpose of the project, email address from an IPCSR institution, etc.); and verify that the data can be opened as expected. If applicable, provide a link to the page where you registered to access the data.

Describe in detail any restrictions on data access and data-sharing, as well as any additional terms of data use that will be relevant for the replication study and final report (e.g. citations that will need to be made). If you were able to access the data because of special permissions that you have, but that you expect other researchers might not have, please document those as well.

The Global Entrepreneurship Monitor's Adult Population Survey (APS) provides the data needed to construct the dependent variable. From the GEM website: "Each year, following data collection, GEM publishes around 20 of its APS indicators and 13 of its NES indicators for all participating economies, via its Global Report and its website. This is an excellent resource for those wishing to learn more about national and global entrepreneurship, and is made freely available to all. However, access to full datasets, which include all GEM indicators, and complete individual-level data for all APS and NES respondents, is restricted to members of GEM National Teams. The National Teams fund and conduct the data collection and are therefore granted a period of exclusive access. These full datasets are only made available to the public 3 years after data collection."

The specific GEM files (downloaded 7/1/2020) are as follows:

- GEM 2001 APS Global Individual Level Data [GEM 2001 APS Global Individual Level Data.sav]
- GEM 2002 APS Global Individual Level Data [GEM 2002 APS Global Individual Level Data.sav]
- GEM 2003 APS Global Individual Level Data [GEM 2003 APS Global Individual Level Data.sav]
- GEM 2004 APS Global Individual Level Data [GEM 2004 APS Global Individual Level Data.sav]
- GEM 2005 APS Global Individual Level Data [GEM 2005 APS Global Individual Level Data.sav]
- GEM 2006 APS Global Individual Level Data [GEM 2006 APS Global Individual Level Data.sav]
- GEM 2007 APS Global Individual Level Data [GEM 2007 APS Global Individual Level Data.sav]
- GEM 2008 APS Global Individual Level Data [GEM 2008 APS Global Individual Level Data.sav]
- GEM 2009 APS Global Individual Level Data [GEM 2009 APS Global Individual Level Data.sav]
- GEM 2010 APS Global Individual Level Data [GEM 2010 APS Global Individual Level Data.sav]
- GEM 2011 APS Global Individual Level Data [GEM 2011 APS Global Individual Level Data_1Feb2015.sav]
- GEM 2012 APS Global Individual Level Data [GEM 2012 APS Global – Individual Level Data_1Feb2015.sav]
- GEM 2014 APS Global Individual Level Data [GEM 2014 APS Global - Individual Level Data_9Mar.sav]
- GEM 2015 APS Global Individual Level Data [GEM 2015 APS Global Individual Data.sav]
- GEM 2016 APS Global Individual Level Data [GEM 2016 APS Global - Individual Level Data.sav]

The file listed for 'GEM 2013 APS Global Individual Level Data' on the GEM website is unable to be opened. Staff of the GEM provided a usable file [GEM 2013 APS Global Individual Level Data.sav] on 5/25/2020 by email upon request.

Some years of the GEM data (1998-2012) are also available through the ICPSR. To ensure all files are handled consistently across years, only the data directly downloaded from the GEM site is used in this replication. However, the codebook available from the ICPSR website was consulted when country code information was ambiguous.

Census data was collected from the Census Bureau's International Data Base. The full IDB dataset can be downloaded from the FAQ page. Per the codebook accompanying the full dataset, the file needed is 'IDBext194.txt,' which contains midyear population by age and sex for each country.

(c) Variable Availability

RR TEAM INSTRUCTIONS: *For each variable required for the replication analysis (listed above), describe the variables from the replication data that can be used to measure it (including which data files or sources each measure is found in),* ***any notes a data analyst should consider when using the measure in a replication analysis****, and any important differences between the original variable and the proposed replication variable.*

If there are multiple variables in the replication data that correspond to a required variable (e.g. two different measures of education in the replication data), include all of those options below. If a variable from the original study ***cannot*** *be measured using the replication data, please make that clear as well.* ***Finally, include a description of the identifiers used to merge multiple datasets, if applicable.***

Entrepreneurship rate

- As mentioned above, the specific item used to measure the entrepreneurship rate is 'manages and owns a business that is up to 42 months old and pays wages.' Though it does not appear to be mentioned explicitly in the paper, this country-year variable appears to be based on respondents aged 20-64 (thus matching the age range of the focal independent variable below).
- For each year of the GEM, the same set of variables are needed in order to create the entrepreneurship rate variable and facilitate its merge with other variables:
    - Year of the survey (yrsurv in all GEM datasets)
    - Country of the survey (country in all GEM datasets)
    - Whether the respondent owns and manages a business that is up to 42 months old and pays wages (babybuso in all GEM surveys prior to 2011; BABYBUSO in GEM surveys after 2010)

Median age

- As mentioned above and in footnote 30 of the paper, the median age is calculated only for those 20–64 years old. The following variables are needed from the Census Bureau's IDB data in order to create the median age variable and merge it with the entrepreneurship rate:
    - Country [Federal Information Processing Standard (FIPS) country/area code]
    - Year
    - Sex [males and females split across each country-year combination]
    - Population at age [...]
        - Specifically, need ages 20-64 for the replication
- To convert the FIPS country codes to country names, the 'codelist' data frame from the countrycode R package was used. As mentioned in the data-processing steps below, the codelist data frame was supplemented with a small handful of country codes that were not already present in the data frame.

Weight

- As mentioned in the Table 2 footnote, "Observations are weighted by the number of individuals who make up each country-year cell." The replication dataset below contains a variable (cy_cell) that counts the number of respondents from the GEM data corresponding to each country-year observation.
    - **Note:** This corresponds to the number of respondents whose data was used to construct the entrepreneurship measure, so it only counts respondents in the age range 20-64.

Year

- The year the GEM survey was administered is included as a variable in the replication dataset.

Country

- The replication dataset below contains two variables that could be used to cluster standard errors at the country level: country or phone_code. The two variables have a one-to-one relationship, so either will work for clustering purposes.

(d) Data Creation

RR TEAM INSTRUCTIONS: *Create a dataset using the data sources and variables listed above. Provide a detailed narrative describing how the various datasets were cleaned and merged into a final replication dataset. Provide a view-only link to a clearly commented script on the OSF that produces the replication data as described in the narrative. Our preference is that this be either an R script or a script from another language that similarly allows for open and reproducible analyses. Please let the SCORE team know if this is not possible.*

- *If the data can be freely shared and posted to OSF, please post it in your OSF project and provide a link to the completed dataset below.*

- *If any part of the dataset cannot be shared between researchers or posted to the OSF, please leave the final dataset off the OSF. Instead, include either below or in your script (commented out at the bottom) two pieces of information that will help an independent team verify they have created the dataset according to your instructions:*
    - *The dimensions of the final dataset(s) you've created (# of rows, # of columns)*
    - *A summary of 8-10 variables in the replication dataset. For numeric variables, the summary should include the mean, standard deviation, and count of NAs. For categorical variables, the summary should include each level present in the data and its count, as well as a count of NAs. If multiple datasets are submitted as part of your work, at least one variable should be included from each dataset.*

The data from the replication sources should be preserved in as 'raw' a form as possible, in order to give the data analyst the most latitude to clean the variables as they see fit. Variables from the original source should be preserved in their original form (e.g. do not recode values of 99 to NA). New variables should only be created when they're needed to complete the merge or combine the datasets; in those cases, please preserve a version of the original, unaltered variable in the new dataset.

Please also use this section to describe:

- *Any deviations between the original study design and the replication design that would result from using this replication dataset.*
- *Any notes about using these variables that you would like to pass along to the data analyst.*

To create the replication dataset, first load the input files (found here and here):

- 16 GEM files listed above, corresponding to the 16 years of GEM data included in the replication (2001-2016)
- The full IDB census dataset (IDBext194.txt)
- The authors' original data (see note in that section for rationale)

GEM data

For each year of the GEM data:

- Filter to ages 20-64 and then compute:
    - The mean entrepreneurship level within each country-year using the specific entrepreneurship variable selected (either babybuso or BABYBUSO depending on year of the GEM)
    - The number of respondents within each country-year
- Within each year of the GEM, keep the variables for mean entrepreneurship level, country-year respondent count, year, country, and country-year

Rename a specific set of country and country-years in the GEM data to facilitate later merging:

- GEM 2012: Rename the Gaza Strip & West Bank to West Bank & Gaza Strip
- GEM 2010: Rename Trinidad & Tobago to Trinidad and Tobago

- GEM 2001, 2002, 2008, 2009, 2010: Rename Korea to South Korea
- GEM 2007: Rename Kazakstan to Kazakhstan
- GEM 2011: Rename the NA value corresponding to phone code 582 to Venezuela, since that's the country associated with 582 for each of the other GEM datasets it appears in.

Combine all years of the GEM data into a single dataset (gem_full)

Census Data

Assign variable names to the Census data (per the Census codebook here), then merge the list of country names from the countrycode data frame into the IDB Census data, and then supply a handful of country names that weren't included:

- GZ = Gaza strip
- NN = Sint Maarten
- OD = South Sudan
- RN = Saint Martin
- UC = Curacao

Rename a set of country names to facilitate merging with GEM data:

- Rename Hong Kong SAR China to Hong Kong
- Rename Czechia to Czech Republic
- Rename Bosnia & Herzegovina to Bosnia and Herzegovina
- Rename North Macedonia to Macedonia
- Rename Palestinian Territories to West Bank & Gaza Strip
- Rename Trinidad & Tobago to Trinidad and Tobago

Limit the Census data to the ages needed to calculate median age (20-64), then sum across male/female to arrive at a total number of cases per age group per country-year. Finally, add in the country names and generate a country-year variable.

For each country-year in the Census data that also appears in the GEM data (gem_full), calculate the median age for those between the ages 20-64 using the function supplied in the R script below.

Merge the GEM data (gem_full) with the median age variable computed from the Census data to create the replication data (replication_data). This dataset has been uploaded to the OSF as replication_data_mkk9.csv.

The R code that constructs the replication dataset according to the procedure above is found here. Please consult this file for details on its use.

Data validation

Note: For purposes of confirming that the GEM and IDB data have been accessed and cleaned in the correct way, the authors' original data is included in the data preparation script. Original data was gathered from the Journal of Political Economy website on 7/15/20 (materials are gated). Per the do file corresponding to the original analysis, the focal analysis being replicated relies on the 'GEM_Country_Year.dta' file.

To validate that the procedures above create the data according to the original authors' procedures, overlapping observations in data (i.e. country-years from 2001-2010) were compared on three focal variables: entrepreneurship rate, median age, and country-year count of GEM respondents.

As documented below, the vast majority of observations have equivalent values, but there are a handful of differences, especially in the variable that was derived from the IDB data (median age). This is most likely because the IDB data has been updated since the original study was published. Because of this, it is recommended that the data analyst relies entirely on the 'raw' data from the GEM and IDB data when conducting the analysis, rather than relying on the authors' data when years overlap.

The following are the discrepancies between the original authors' data and the re-constructed data that overlaps with the original data (i.e. years 2001-2010):

- For the median age variable (calculated from the IDB Census data), 360 observations have the same value in the two datasets (this is, no discrepancy), 27 observations are off by 1, 3 observations are off by 2, and 3 observations are off by 3.
    - The observations that are different between the two data sources tend to be concentrated in the same country (e.g. Jamaica_2005, Jamaica_2006, Jamaica_2008, Jamaica_2009, Jamaica_2010 are five of the country-years that don't match up exactly).
- For the entrepreneurship variable (calculated from the GEM data), only one overlapping country-year is off by more than .001 (United Kingdom_2005).
- For the country-year respondent count variable (calculated from the GEM data and used for weighting), only one overlapping country-year has a different respondent count (United Kingdom_2005, as above).

Finally, there are nine country-years appearing in the replication data from 2001-2010 that do not appear in the original data.

- The first three can be safely dropped from the analysis for the following reasons:
    - Puerto Rico_2007
        - IDB Census data is not available for Puerto Rico before 2010, so this has no value for the 'median age' variable.
    - Shenzhen*_2009
        - This city does not appear in the IDB Census data and thus has no value for the 'median age' variable.

    - * 'Azores'_2010
        - This does not appear in the IDB Census data and thus has no value for the ‘median age’ variable. The data finder (A. Tyner) recommends that this be dropped from the analysis, though it could be incorporated as part of the analysis of the Portugal_2010 observation (which is already part of the replication data).
- The remaining six country-years below all have Census data available, and thus can be included in the replication analysis. It is not clear why they were not included in the original analysis:
    - Australia_2001
    - Portugal_2004
    - New Zealand_2005
    - Hong Kong_2009
    - West Bank & Gaza Strip_2009
    - West Bank & Gaza Strip_2010

(e) Data Dictionary

RR TEAM INSTRUCTIONS: *Create a data dictionary following this template. Provide below a view-only link to the completed data dictionary included in the OSF project. If the Data Analyst will need to create new variables using the variables in the final replication dataset (e.g. recoding the provided education variable to be in a better format for analysis), please document below your recommendation on how the analyst should do so. Please also document any additional notes regarding the variables in the dataset that do not fit within the provided data dictionary template or the other sections above.*

The data dictionary for the replication dataset (replication_data_mkk9.csv) is found on the OSF here.

13. Sample size

RR TEAM INSTRUCTIONS: *Please report below the analytic sample size(s) in the replication dataset, with reference to however many units or levels are in the data. Please report as much information here as will be helpful for the review committee to be aware of, including differences in sample size resulting from various analytic decisions (e.g. listwise deletion vs multiple imputation).* ***Finally, when the replication combines observations from the original study with new observations, please estimate what proportion of the analytic sample’s observations will be comprised of original vs. new observations.***

Data finders’ response goes here: The full replication dataset has 789 observations, though three observations (Puerto Rico_2007, Shenzhen*_2009, and * 'Azores'_2010) have no median age variable and should probably be dropped from the analysis. Of the remaining 786 observations, 402 are from 2001-2010 and 393 of those are the same observations that were included in the original analysis.

Based on the above paragraph, the number of new observations will be between 384-393 depending on whether the data analyst retains any of the nine observations from 2001-2010 that appear in the replication data but not in the original data.

Required sample size [to be filled out by the SCORE team]: For the *sample size calculations*, the primary unit of analysis is the **country**. An estimate of the minimum viable sample size for the data analytic replication is: 5. For comparison, the stage1 required sample size would be: 22 and the stage2 sample size would be: 50.

Note: SE are clustered within countries, so the power analysis currently assumes the same data structure (in terms of years per country) between the original and replication. So, *to get the total observations needed, multiply the Ns above by the years per country in the original* (which looks to be about 7 on average). If the total N is arrived at by using fewer years per country but more countries, this should result in higher than expected power, assuming the same correlation/ICC structure in the original and replication.

14. Sample size rationale

For data analytic replications in SCORE, three sample sizes are calculated:

- *A minimum threshold sample size, defined as the sample size required for 50% power of 100% of the original effect*
- *A stage 1 sample size, defined as the sample size needed to have 90% power to detect 75% of the original effect*
- *A stage 2 sample size, defined as the sample size needed to have 90% power to detect 50% of the original effect*

Details about how those sample sizes were calculated for this project are found here.

15. Stopping rule (provided by SCORE)

RR TEAM INSTRUCTIONS: *Because all existing data replications that clear SCORE's minimum power threshold will proceed to analysis, the stopping rule is not relevant for these kinds of projects.*

N/A -- all observations will be used in a single analysis.

Variables

RR TEAM INSTRUCTIONS: *The preregistration form divides variables across three questions: manipulated variables, measured variables, and indices (i.e. analytic variables derived from raw variables). For existing data replications, only fill out the "Measured variables' and 'Indices' sections. Please do not fill out anything in the 'Manipulated variables' section.*

The raw data of any transformed variable (e.g. reaction time → log reaction time) or any created index should be defined in the 'Measured variables' section. Details regarding the variable transformation should be specified in the 'Transformations' section. Details regarding the creation of an index should be specified in the 'Indices' section.

Across these questions, you should define all variables that will later be used during your analysis (including data preparation/processing). You can describe all variables in the preregistration and/or summarize and link to a data dictionary (codebook) in your repository to answer these questions.

If you will share data from your replication, this is also the place to state whether any variables will be removed prior to sharing the dataset (e.g. to reduce risk of participant identification or comply with copyright restrictions on scale items.)

16. Manipulated variables

RR TEAM INSTRUCTIONS: *Manipulated variables in this preregistration refer specifically to variables that have been randomly assigned in an experiment. The use of data from an experiment should be rare in existing data replications. If your existing data replication relies on experimental data, please document each manipulated variable as a measured variable, and use the codebook to indicate what each level of the variable corresponds to (e.g. participants assigned to the treatment condition = 1; participants assigned to the control condition = 0). The default language in bold below has been copied into all existing data replication preregistrations.*

N/A -- not documented for existing data replications.

17. Measured variables

RR TEAM INSTRUCTIONS: *Please use this section to document each variable that was used in the original study's analysis and the role it served (e.g. dependent variable, control variable, sample parameter, etc). For each variable, provide the description of the variable offered in the paper and/or codebook of the original study, the variable in the replication dataset that it corresponds to, and explain any deviations between the two. In cases where an equivalent*

replication variable was not found, explain how, if at all, you expect it will affect the replication attempt. In cases where you are adding a variable that was not present in the original study, please explicitly state that you are doing so, and explain how, if at all, you expect it will affect the replication attempt.

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration surface all of the variables needed to replicate the focal analysis?*
- *Are deviations between the original variables and replication variables documented when needed?*

Entrepreneurship Rate

- This is the dependent variable in the original model.
- Entrepreneurship rate is defined as the proportion of the labor force that owns a business that is up to 42 months old. This data is taken from the Global Entrepreneurship Monitor (GEM) from 2001 to 2010. In the csv file, this is listed as "entrepreneurship"

Median Age of Cohort (a)

- The independent variable that serves as the main focal point of the original study.
- The variable a is a measure of how comparatively young or old one country's workforce is compared to another, using the median age of the cohort as an evaluation of center. This is specifically restricted to the ages of 20 to 65 as 65 is the typical retirement age and 20 is the typical age where individuals join the workforce.

Year

- Dummy variables for the year the data points are taken from. These are used in accordance with the original study.

18. Indices

RR TEAM INSTRUCTIONS: *If any of the measured variables described in Section 17 will be combined into a composite measure (including simply a mean), describe in detail what measures you will use and how they will be combined. Please be sure this preregistration includes a link to a clearly commented script that constructs the index according to the narrative.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify each of the composite measures (e.g. mean scores, factor scores) that are needed for the focal analysis, and which of the measured variables in Section 17 are used in each one (e.g. the happiness, joy, and satisfaction items will be used to create the 'positive feelings' measure)?*
- *Does the preregistration link to a clearly commented script that constructs the indices according to the narrative description?*

Population and Country Weights

- Given observations in the dataset are stratified by country and year, they are each weighted by the population of the country. This is done by using the “cy_cell” variable as an analytic weight in the regression, just like the original study.
- Standard errors are similarly weighted by clustering observations based on the “country” variable, just like the original study.

Analysis Plan

19. Statistical models

RR TEAM INSTRUCTIONS: *This section should describe in detail the analysis that will be performed to replicate the focal result. This analysis must align as closely as possible with the original study’s analysis, even if you have identified limitations in the original study. The level of detail should allow anyone to reproduce your analyses from your description below. Examples of what should be specified: the model; each variable; adjustments made to the standard errors and to case weighting; additional analyses that are required to set up the focal analysis; and the software used.*

Beyond the replication of the focal analysis from the original study, it is at your discretion to test the claim using other analytic approaches as a check of the robustness of the claim. The original test should be listed first and be clearly distinguished from any other tests. If you are testing additional confirmatory hypotheses, describe them in the same order as you numbered them in the “Hypotheses” section above and make clear reference to the specific hypothesis being tested for each.

Please provide a link to a clearly commented script that performs the analysis described in the narrative provided below. Our preference is that this be either an R script or a script from another language that similarly allows for open and reproducible analyses. Please let the SCORE team know if this is not possible. Please also test that the code runs without error on a random subset of 5% of the replication dataset, and provide verification that the code has produced a sensible result below (a screenshot of the results is preferable). Finally, please confirm that you have only developed and tested your analysis plan and code using 5% of the data.

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify which statistical model will be used to provide the ‘focal evidence’ for the SCORE test (e.g. a regression coefficient in a larger multiple regression model), and does it correspond closely to the model and evidence from the original study?*

- *Does the preregistration describe each variable that will be included in the focal analysis, and what role each variable has (e.g. dependent variable, independent variable)?*
- *Does the preregistration include a detailed specification of the focal analysis, including interactions, lagged terms, controls, etc., in both narrative form and in a clearly commented script?*
- *Does the preregistration verify that the code runs without error on a random subset of the replication dataset?*

The model in the original study states that the cohort structure of a given country is:

$$f(a,r) = \frac{r}{e^r-1}e^{ra}$$

With a being age and r being the cohort shrink rate. The cdf of age is:

$$F(a,r) = \frac{e^{ra}-1}{e^r-1}$$

The relationship between the age of a country's labor force and the rate of entrepreneurship is considered by running a nonlinear least squares regression to estimate the cohort shrink rate and factoring it into the regressions on the various measures of a work force's age. In this replication we focus on the median age of a country's workforce as the relevant measure of a and see its relationship with a country's entrepreneurship rate in any given year.
[Statement of disclosure relating to the selection in section 10 above]: Note that, in troubleshooting the code for the regression, it was run on the entire dataset before creating the 5% representative samples in the OSF project. However, the regression containing the full data is not referenced in the OSF project and the plan of analysis was established prior to this being run. The .do files for 5% of the entire dataset (https://osf.io/ryqut/?view_only=6c87f474dc594070817403ed417a041d), 5% of the old data (https://osf.io/xjzfn/?view_only=6c87f474dc594070817403ed417a041d), and 5% of the new data (https://osf.io/fx4mc/?view_only=6c87f474dc594070817403ed417a041d) can be viewed in the OSF project along with a .csv file containing the coefficients and standard deviations from each(https://osf.io/s4n6c/?view_only=6c87f474dc594070817403ed417a041d,https://osf.io/vbn73/?view_only=6c87f474dc594070817403ed417a041d,https://osf.io/5sz87/?view_only=6c87f474dc594070817403ed417a041d). Note also that these .do files are written to work on a MacOS system (specifically Victor Volkman's Macbook pro) and directories may need to be changed to run them on a Windows operating system.

20. Transformations

RR TEAM INSTRUCTIONS: *This section should describe how any of the measured variables or composite measures mentioned above will be transformed prior to the analyses listed in Section 19. These are adjustments made to variables **after** measurement or measure creation, and might include centering, logging, lagging, rescaling etc. Please provide enough detail such that anyone else could reproduce the transformations based on the description below. Please be sure this preregistration includes a link to a clearly commented script that performs the transformations described in the narrative provided below.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration specify which of the measured variables or composite measures will need to be transformed prior to the focal analysis?*
- *For each variable needing transformation, does the preregistration adequately describe the transformations, including any centering, logging, lagging, recoding, or implementation of a coding scheme for categorical variables?*
- *Does the preregistration link to a clearly commented script that performs each transformation?*

In our regressions, the relevant independent and dependent variables will not be transformed beyond weighting them based on country and population. Data will only be transformed to recover the cohort shrink rate r in the vein of the original paper. This would include normalizing a=0 to represent age 20 and a = 1 to represent age 65.

21. Inference criteria

RR TEAM INSTRUCTIONS: *This section describes the precise criteria that will be used to assess whether the hypotheses listed above were confirmed by the analyses in Section 19. The default language below only applies to the test of the SCORE claim, **H***. It is at your discretion to describe the inferential criteria you will use for any additional analyses. They need not rely on p-values and/or the same alpha level we have specified for **H***.*

If the additional analyses will use multiple comparisons, the inference criteria is a question with few "wrong" answers. In other words, transparency is more important than any specific method of controlling the false discovery rate or false error rate. One may state an intention to report all tests conducted or one may conduct a specific correction procedure; either strategy is acceptable.

Criteria for a successful replication attempt for the SCORE project is a statistically significant effect (alpha = .05, two tailed) in the same direction/pattern as the original study on the focal hypothesis test (**H***). For this study, this criteria is met if: (a.) the coefficient of entrepreneurship is proven to be statistically significant and (b.) the coefficient is negative as shown in the original paper. Special attention may also be given to the difference between the coefficient found in the original paper and those of our replication, taking newer data into account.

22. Data exclusion

RR TEAM INSTRUCTIONS: *The section below should describe the rules you will follow to exclude collected cases from the analyses described in Section 19. Note that this refers to exclusions **after** the creation of the replication dataset; exclusion criteria that prevent a case*

from entering the replication dataset in the first place should be detailed in the 'Data Collection Procedure' section above. Please be as detailed as possible in describing the rules you will follow (e.g. What is the specific definition of outliers you will use? Exactly how many attention checks does a participant need to fail before their removal from the analytic sample?).

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration comment on whether any cases included in the replication dataset will be excluded prior to data analysis?*
- *If yes, does the preregistration provided detailed instructions on how the exclusions will be performed (e.g. Is the definition of outlier provided? Is the number of attention checks failed before a participant is excluded specified?)*

As in the original study, we will only be considering workers from the ages of 20-65. This is justified in the original paper by 65 being the usual retirement age and 20 being the usual age for individuals to enter the workforce. However, this does not become an issue with the dataset being used in this replication as the "median_age" variable never violates these parameters.

23. Missing data

RR TEAM INSTRUCTIONS: *The section below should describe how missing or incomplete data will be handled. Please be as detailed as possible in describing the exact procedures you will follow (e.g. last value carried forward; mean imputation) and any software required (e.g. We will use Amelia II in R to perform the imputation).*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration comment on how missing or incomplete data will be addressed (e.g. casewise removal, missing data imputation)?*
- *If applicable, does the preregistration specify how many missing variables will lead to a case's removal (e.g. If a subject does not complete any of the three indices of tastiness, that subject will not be included in the analysis.)?*
- *If applicable, does the preregistration describe how missing data imputation will be performed, including relevant software?*

Any year-country cells with no median age variable such as Shenzhen 2009 and Puerto Rico 2007 were dropped prior to creating the population weights or running regressions. This caused our dataset to be limited from 790 observations to 786.

24. Exploratory analysis (Optional)

RR TEAM INSTRUCTIONS: *If you plan to explore your data set to look for unexpected differences or relationships, you may describe those tests here. An exploratory test is any test where a prediction is not made up front, or there are multiple possible tests that you are going to*

use. A statistically significant finding in an exploratory test is a great way to form a new confirmatory hypothesis, which could be registered at a later time. If any exploratory analyses involve additions to the data collection procedure beyond what was performed in the original study (e.g. additional items on the survey; running another condition in the experiment), please describe them below.

25. Other

RR TEAM INSTRUCTIONS: *This section serves two purposes. First, please use this section to discuss any features of your replication plan that are not discussed elsewhere. Literature cited, disclosures of any related work such as replications or work that uses the same data, plans to make your data and materials public, or other context that will be helpful for future readers would be appropriate here. Second, please also re-surface any major deviations from earlier in the preregistration that you expect a reasonable reviewer could flag for concern. Give a summary of these deviations, focusing on larger changes and any possible challenges for comparing the results of the original and replication study.*

Specific points to keep in mind (please also consult the Reviewer Criteria):

- *Does the preregistration reference other sections of the preregistration where substantial deviations from the original study have been described (including deviations due to differences in location or time compared to the original study)?*
- *Does the preregistration comment on plans to make the data and materials from the replication study public?*

Final review checklist

REVIEWER INSTRUCTIONS: *For the following questions, reviewers please indicate whether you can 'sign off' on the following items by adding a comment. You can update this response as the lab moves through revisions during the review period!*

- Included in this pre-registration are specific materials needed to create a replication dataset:
    - Is the final replication dataset that the research team constructed suitable for performing a high-quality, good-faith replication of the focal claim selected from the original study?
    - Is the procedure for constructing the final replication dataset sufficiently documented that an independent researcher could construct the same dataset following the procedures and code they lay out?
- Included with this pre-registration is a narrative description of how the replication dataset will be used to perform the focal replication analysis, as well as the specific analytic scripts/code/syntax that will be used:
    - Is the analysis plan (including code) that's documented in the preregistration consistent with a high-quality, good-faith replication of the focal claim selected from the original study?
    - Has the data analyst demonstrated that the analysis code works as expected on a random 5% of the final replication dataset?
- I have reviewed all sections of this pre-registration, and I believe it represents a good-faith replication attempt of the original focal claim.