Replication of a Research Claim from Hossain (2020),
from medRxiv

Replication Team: Carolin Nast and Esteban Méndez-Chacón

Research Scientist: Nick Fox

Action Editor: Kevin Esterling

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

Reviewer #1: [Dino Krupić] DONE

Reviewer #2: [Elisabeth Julie Vargo]DONE

Reviewer #3: [Josh Metacotta]

Review Period: August 26 - August 31

View-only links to: Original Paper, 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 demonstrate that the focal analysis produces sensible results using just that random sample (see section 19 for details). **Do not use the rest of the data until after your study is registered and 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 up to three analyses (details immediately below). This will likely require you to subset your data, 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 relies only on the ‘new data’ (the focal analysis), one analysis that relies on all available data, and a third analysis that relies only on the original data. Please make sure all three 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 Hossain_covid_b344

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 Hossain (2020).

2. Authors and affiliations

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

Carolin Nast A[1]
Esteban Méndez-Chacón [2]

1 Student, Utrecht University
2 Central Bank of Costa Rica

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 Hossain (2020) is that a cohesive society where people are involved in more festivals and religious gatherings is more likely to have more cases of COVID-19.This reflects the following statements from the paper's abstract: The estimation results indicate that the number of confirmed cases of Coronavirus infection is higher in countries with lower yearly average temperature, higher economic openness and stronger political democracy. The description of the analysis is as follows: ‘By estimating a regression equation where the dependent variable is the total number of cases confirmed infection per one million people in a country on a day (03 April 2020) and the predictor variables are the democracy index of the country (the proxy variable for social cohesiveness), the yearly average temperature of the country, and the openness of the country (measured by international trade as a percentage of GDP)...We apply Least Squares method on model (1) and find that precipitation and population density have no significant effect on the number of infection cases per one million people (Y). Those variables are then excluded, and the model is re-estimated. In the re-estimated model, the variables average temperature, openness and democracy appear as highly significant.’ The focal finding is: ‘The positive sign of democracy index indicates that more democratic countries are affected more by the disease. From Table 1: model 1 coefficient = 86.76467, p = 0.0001.'

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*: At the country level, the democracy index will be positively associated with the total number of confirmed infections per one million people.

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 data finders 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 original study is investigating the relationship between the severity of the COVID-19 disease and the environmental, economic and social factors of countries. The data used for the analysis are publicly available. The study combines different data sources:

1. Coronavirus infection cases along with population in 2018 are collected from the website of European Centre for Disease Prevention and Control (ECDC).
2. Average yearly temperature data were collected from meteoblue.com.
3. Average precipitation data were collected from indexmundi.com.
4. Openness data as defined by ratio of international trade to GDP from Databank of the World Bank.
5. Scores of democracy index 2019, compiled by the Economist Intelligence Unit, UK, is collected from Wikipedia.
6. Population density (data source is not mentioned).

The dependent variable, severity of the COVID-19 disease, is calculated by counting the infection cases for all countries from 31 December 2019 until 03 April 2020. This number of total cases per country is converted to cases per one million population. The focal analysis for replication does not include the independent variables of population density and average precipitation.

From the original study: “Total number of cases of Coronavirus infection by countries is time variant. The number accumulates over time. Nonetheless, total cases of infection in a specific country on a specific day (Y) is a fixed number. Thus, the variable, total cases of infection by countries has both time series and cross-section elements. We collect data on total infection cases by countries from 31 December 2019 to 03 April 2020. Total cases of infection are converted to cases per one million population to capture the population effect. Cases of infection per one million people on 03 April 2020 by countries are denoted by Y and used as the dependent variable in our experimentation. Values of this variable obviously depend on its previous values. Consequently, Y is simultaneously a cross-section and time series variable.” (p. 10).

The author states that he cannot include all countries infected by COVID-19 in the analysis due to unavailability of all data for few countries. In total, 163 countries are included in this analysis (p. 10). It is not mentioned which countries specifically.

The replication dataset contains three versions of the COVID variable. One version covering the total dates available (December 31 2019 - August 11 2020). A second version covers the dates from the original study (December 31 2019 - April 03 2020). A third version covers the dates that have occurred since the original study (April 04 - August 11). Similar to the original study, all independent variables are included in the replication dataset.

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.*

No randomization was applied to the data finders knowledge.

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

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 raw datasets for replication have been accessed, cleaned and one variable has been manipulated (Openness, see 12c). The final replication dataset includes all variables needed for analysis. The variables in the final dataset correspond to the constructs used in the original study, however, two variables have been collected from a different data source (see 12c). The data analyst has to calculate the final dependent variable, COVID cases per 1 million population.

The final replication dataset can be found here: https://osf.io/6cq9b/

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.*

The unit of analysis of the original study are 163 countries.

Dependent variable:

Coronavirus infection cases

- Datasource: European Centre for Disease Prevention and Control (ECDC)
- The dependent variable, severity of the COVID-19 disease, is calculated by counting the infection cases for all countries from 31 December 2019 until 03 April 2020. This number of total cases per country is converted to cases per one million population.

Independent variables:

Environmental variables

Average yearly temperature

- Yearly average temperature (averaged for the years 1961-1990)
- Data source: meteoblue.com (as indicated in the original paper on p. 9)
- Further note: The original author sent the SCORE team some clarification about this variable and provided the original data. I, the data finder, strongly guess that the author used the yearly average temperature data from Wikipedia and not as indicated in the original paper from the website meteoblue. Please see the screenshot I uploaded to the OSF project (Link: https://osf.io/tkdfs/). The source given by Wikipedia for their data is the Lebanese-economy-forum. Here is the citation (with link) from Wikipedia: "Average yearly temperature (1961-1990, Celsius) - by country". *lebanese-economy-forum.com*. Lebanese Economy Forum. Archived from the original on 5 September 2015. Retrieved 20 August 2015." Another indication for my guess is that I intensively searched on the meteoblue website and could not find this historical weather data the author claims to have used. Due to the questionable reliability of data retrieved from Wikipedia, I have included comparable historical weather data from the World Bank (as the World Bank is a very reliable data source). For further information see Variable Availability.

Economic and social variables

Openness

- Operationalization: Ratio of international trade to GDP
- Datasource: World Bank
- Further note: It is not indicated in the original paper on which year(s) this variable is based.

Democracy index 2019

- Datasource: collected from Wikipedia, however, data is compiled by Economist Intelligence Unit, UK
- Further note: Due to the questionable reliability of data retrieved from Wikipedia, I have included this variable from another data source, namely gapminder. For further information see Variable Availability.

Sample size of analysis has 163 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 data sources for replication are (nearly) the same as in the original study (two variables are collected from a different datasource, see Variable Availability). The data is freely available for research purposes and offers a very user-friendly platform. The data files have been successfully downloaded. No registration is needed to download the necessary data.

For citations in academic journals the following can be used for the four variables:

Coronavirus infection cases

Suggested citation by data finder: “European Centre for Disease Prevention and Control (2020). Data on geographic distribution of COVID-19 cases worldwide. Retrieved from https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide"

Average yearly temperature
Recommended citation from the website: "Derived from the Climate Research Unit (Mitchell et al, 2003). Retrieved from https://datacatalog.worldbank.org/dataset/climate-change-knowledge-portal-historical-data"

Openness
Suggested citation by data finder: "The World Bank. Data on Trade (% of GDP). Retrieved from https://data.worldbank.org/indicator/NE.TRD.GNFS.ZS"

Democracy index 2019
Suggested citation by data finder: "EIU. Democracy index 2019. Retrieved from https://www.gapminder.org/data/documentation/democracy-index/"

(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.***

Dependent variable:

Coronavirus infection cases

- Datasource: European Centre for Disease Prevention and Control (ECDC)
- The confirmed COVID cases are included in three versions in the replication dataset, covering three different time frames (see below).
- The data analyst has to calculate the total cases per country per 1 million population. For this reason population data is also retrieved from ECDC.
- Link: https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

Confirmed COVID cases 12.31.2020 to 04.03.2020 (time frame of the original study)

- Cumulative number of cases calculated based on daily cases between 12.31.2020 and 04.03.2020

Confirmed COVID cases 12.31.2020 to 08.11.2020 (whole available time frame)

- Cumulative number of cases calculated based on daily cases between 12.31.2020 and 08.11.2020

Confirmed COVID cases 04.04.2020 to 08.11.2020 (new data)

- Cumulative number of cases calculated based on daily cases between 04.04.2020 and 08.11.2020

Variable needed for the calculation of COVID cases per 1 million population

Population

- Variable data source: ECDC

Independent variables:

Environmental variables

Average yearly temperature

- Data source of the original study: meteoblue.com
- Data source replication dataset: World Bank
- Attention: To my knowledge (datafinder), this information is not available on meteoblue.com. Instead, the exact same data can be found on wikipedia. Therefore, the replication dataset includes data on the annual temperature based on historical data from the World Bank, averaged for the years 1961 to 1999)
- Link to the data: https://datacatalog.worldbank.org/dataset/climate-change-knowledge-portal-historical-data
- Difference to original study: The original study used data on the annual temperature based on historical data, averaged for the years 1961 to 1990, while the variable included in the replication dataset used annual temperature, averaged for the years 1961 to 1999).

Economic and social variables

Openness

- Operationalization: Ratio of international trade to GDP
- Datasource: World Bank
- Further note: The author does not specify for which year he collected this variable from the World Bank. The World Bank trade data is available from 1960-2019. Due to the fact that a significant number of countries did not have a value for 2019, I decided to include the most recent value within the time frame of 2015 - 2019 (variable trade.recent). This

means that the value is either imputed from 2015, 2016, 2017, 2018, or that the value was available for 2019. I decided not to include older values in order to guarantee validity. To make this imputation transparent, I created a variable called imputed.trade, which indicates if international trade value is from 2019 (0), or was imputed (1) (variable imputed.trade). I also uploaded the original dataset, so that the data analyst can change the above explained imputation if necessary. Additionally, I followed the suggestion of the SCORE team and included another variable (trade.2016) that refers to international trade in 2016 for all available countries. This year was chosen because it is the most recent year for which the most number of countries have valid responses. The data analyst has to decide which variable to use for the openness construct.

- Link to the data: https://data.worldbank.org/indicator/NE.TRD.GNFS.ZS

Democracy index 2019

- Data source of the original study: collected from Wikipedia, however, data is compiled by Economist Intelligence Unit, UK
- Further note: The data can not be directly downloaded from The Economist. The Economist only publishes an annual report (pdf) where they present their Democracy index.
- Data source of the replicatication: Gapminder
- Link to data: https://www.gapminder.org/data/documentation/democracy-index/

Organizational variables

Country name

Country Code

(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.*

The attached uncompiled R-notebook script is very detailed in its sections and can be found here: https://osf.io/6xgeh/

The final replication dataset can be found here: https://osf.io/6cq9b/

The data analyst has to compute the following final dependent variable:

- COVID cases per 1 million population (see Variable Availability)

Additionally, I want to point out the imputation approach explained in Variable Availability concerning the variable Openness again. The data analyst can either follow this approach (most recent international trade data from 2015-2019), use the alternative variable that includes international trade data only for the year 2016, or use the original data provided.

The original Annual Average Temperature data from the World Bank can be found here: https://osf.io/g9mhe/

The original democracy index data from Gapminder can be found here: https://osf.io/x4j7d/

The original data on international trade from the World Bank can be found here: https://osf.io/93gu2/

The original COVID data from European Centre for Disease Prevention and Control has been accessed via the website: https://opendata.ecdc.europa.eu/covid19/casedistribution/csv

(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 can be found here: https://osf.io/u2x6b/

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 unit of analysis of the original study are 163 countries. It is not clear which countries were specifically included in the original analysis. Countries with any missing value for COVID cases, population, temperature, openness, or democracy index, are deleted from the replication dataset. After listwise deletion, the replication dataset includes 150 countries. This difference of 13 countries compared to the original study can potentially be explained by the use of a different data source for the variable Average yearly temperature (world bank instead of wikipedia), and the years used for the imputation of the Openness (international trade) variable (see 12c). Both of these aspects are not stated/explained in the original study. In order to find out the exact reasons, a list of countries included in the original study is necessary.

Required sample size [to be filled out by the SCORE team]: The primary unit of analysis is the country. An estimate of the minimum viable sample size for the data analytic replication is: 38. For comparison, the stage1 required sample size would be: 126 and the stage2 sample size would be: 188.

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: https://osf.io/nkubx/?view_only=c48bded349914d1cae1318aeef0f5ded

15. Stopping rule (provided by SCORE)

RR TEAM INSTRUCTIONS:

There are three potential analyses that could be performed with the provided data, and SCORE recommends performing at least two of them:

- **The focal analysis that relies only on dates not covered by the original study.**
- **An additional analysis that relies on all available dates.**
- **An optional third analysis that relies only on dates covered by the original study.**

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?*

Note: Lines 68 to 112 of the Stata do-file "Hossain 2020 - Replication Analysis 5% random sample.do" generate the variables described below.

CASES_PER_MILLION

- Use in the analysis: Dependent variable.

- Description from the original study: “Total cases of infection are converted to cases per one million population to capture the population effect” (page 10)

The dependent variable “Y is the total number of cases of confirmed infection per one million people in a country on a day (03 April 2020)” (page 11)

While in the original study the time frame is April 03 2020, in the replication it depends on the type of analysis chosen.

- Variables used in the replication: TOTAL_CASES and POPDATA2019.
- CASES_PER_MILLION corresponds to the cases per one million population.
- No deviations between the original study and the replication study.

DEMOCRACY

- Use in the analysis: Focal independent variable.
- Description from the original study: In page 11, the author details the regression estimated in the paper. One of its control variables is the democracy index of countries, that in the replication analysis is the focal independent variable.
- Variables used in the replication: DEMOCRACY.
- DEMOCRACY corresponds to a continuous variable that expresses the quality of democracy as a number between 0 and 10. A higher number means a country with stronger political democracy.
- No deviations between the original study and the replication study. However, it is important to take into consideration that while The Economist publishes the index with a scale from 0 to 10, Gapminder (the data source for democracy index) has converted the index from 0 to 100 (see https://www.gapminder.org/data/documentation/democracy-index/). Because the author in the original study used the scores of democracy index compiled by the Economist Intelligence Unit (page 9), the data obtained in Gapminder is divided by 10.

TEMPERATURE

- Use in the analysis: Control variable.

- Description from the original study: In page 11, the author details the regression estimated in the paper. One of its control variables is the yearly average temperature of countries.

- Variables used in the replication: ANNUAL_TEMP.

- TEMPERATURE corresponds to a continuous variable representing yearly temperature averaged between 1961 and 1999.

- As it is mentioned in Section 12, the original study used data on the annual temperature based on historical data, averaged for the years 1961 to 1990, while the variable included in the replication dataset used annual temperature, averaged for the years 1961 to 1999).

OPENNESS

- Use in the analysis: Control variable.

- Description from the original study: In page 11, the author details the regression estimated in the paper. One of its control variables is openness measured by international trade as a percentage of GDP of countries.

- Variables used in the replication: TRADE_2016.

- OPENNESS corresponds to a continuous variable representing exports of goods and services, as a percentage of GDP.

- As it is explained in Section 12, it is not clear the specific year for which the author collected the openness variable. In the replication it is used international trade in 2016. This year was chosen because it is the most recent year for which the most number of countries have valid responses.

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?*

CASES_PER_MILLION, the dependent variable. It is constructed using the total COVID cases (TOTAL_CASES) and population (POPDATA2019) variables, according to the formula (TOTAL_CASES /POPDATA2019)*1000000 . Lines 68 to 89 of the Stata do-file "Hossain 2020 - Replication Analysis 5% random sample.do" construct the CASES_PER_MILLION variable.

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?*

This statement confirms that only 5% of the data have been randomly sampled in developing the analysis plan and code contained in this preregistration.

Lines 19 to 66 of the Stata do-file "Hossain 2020 - Replication Analysis 5% random sample.do" generate a random sample that contains only 5% of the data. Before the sample is drawn, the type of data that is used in the analysis needs to be indicated through the local "dataset":

1. A dataset value equal to 1 draws the random sample from the data used in the original study. This is, cases of infection per one million people on April 03 2020.

2. A dataset value equal to 2 draws the random sample from the data after the original study was conducted. This is, cases of infection per one million people after April 03 2020.

3. A dataset value equal to 3 draws the random sample from the whole-time frame. This is, cases of infection per one million people between December 31 2019 and August 11 2020.

4. A dataset value equal to 4 draws the random sample from the time frames not used in the original study. This is, cases of infection per one million people between December 31 2019 and August 11 2020; and cases of infection per one million people after April 03 2020.

5. A dataset value equal to 5 draws the random sample from all available time frames.

To replicate the analysis an ordinary least squares (OLS) regression is estimated. The dependent variable indicates the total number of cases of confirmed infection per one million people in a country on a day. The focal independent variable is the democracy index. Moreover, temperature and openness are included as covariates. The software used is Stata 15.1

The analysis code is "Hossain 2020 - Replication Analysis 5% random sample.do" A log file with the Stata session demonstrates that this code works:

1. File “Hossain_5_Random_Sample_Original_Study.pdf” shows how the analysis code works using a random 5% of the original study data. This is, cases of infection per one million people on April 03 2020.

2. File “Hossain_5_Random_Sample_After_Original_Study.pdf” shows how the analysis code works using a random 5% of the data after the original study was conducted. This is, cases of infection per one million people after April 03 2020.

3. File “Hossain_5_Random_Sample_Whole_Time_Frame.pdf” shows how the analysis code works using a random 5% of the data of the whole-time frame. This is, cases of infection per one million people between December 31 2019 and August 11 2020.

4. File “Hossain_5_Random_Sample_Not_Original_Study.pdf” shows how the analysis code works using a random 5% of the time frames not used in the original study. This is, cases of infection per one million people between December 31 2019 and August 11 2020; and cases of infection per one million people after April 03 2020.

5. File “Hossain_5_Random_Sample_All_Available_Time_Frames.pdf” shows how the analysis code works using a random 5% of the all available time frames.

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?*

The variable DEMOCRACY is divided by 10 to obtain an index with a scale from 0 to 10. The reason is that while The Economist publishes the index with a scale from 0 to 10, Gapminder (the data source for democracy index) has converted the index from 0 to 100. Because the author in the original study used the scores of democracy index compiled by the Economist Intelligence Unit (page 9), the data obtained in Gapminder is divided by 10.

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 pattern as the original study on the focal hypothesis test (**H***).

To test the SCORE claim, H*, that at the country level, the democracy index will be positively associated with the total number of confirmed infections per one million people, an Ordinary least squares (OLS) regression is estimated. The specification takes the form:

$$CASES\_PER\_MILLION_i = \delta + \gamma DEMOCRACY_i + \beta X_i$$

Where:

1. $CASES\_PER\_MILLION_i$: is the total number of cases of confirmed infection per one million people in country i on a day. It is the dependent variable.
2. $DEMOCRACY_i$: a continuous variable ranging from 0 to 10, and where higher values represent a stronger political democracy. It is the focal independent variable. The coefficient of interest is γ , which indicates how the quality of democracy is associated with the total number of confirmed infections per million people.
3. δ : constant term
4. X_i : vector of control variables for country i . According to page 13, the covariates in Model 1 of Table 1 are:
    1. TEMPERATURE
    2. OPENNESS
5. β : vector of coefficients for X_i

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?)*

No observations are excluded from the analysis.

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?*

The replication data do not contain any missing data.

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.