Universidad de los Andes
Facultad de Economía

ECONOMIA URBANA

2022-1

Econ 4683

Ignacio Sarmiento-Barbieri

Deliverable 2: Assessment

Due: March 18th at 6 pm on Bloque Neón

The second deliverable will focus on the assessment stage of the paper's reproduction, and as with every deliverable we will lean on the ACRe Guidelines

This deliverable has two parts. In the first parts, you will describe the inputs for your reproduction. Based on this, in the second part you will assign a Reproducibility Score.

1 Description of the inputs

1.1 Description of the data sources and raw data

In the Assessment stage, you will describe the available reproduction materials and assign a reproducibility score to the display items associated with your selected claims. You will also review reproducibility practices for the overall paper.

In the paper you chose, find references to all data sources used in the analysis. A data source is usually described in narrative form. For example, if in the body of the paper, or the appendix, you see text like "(...) for earnings in 2018 we use the Current Population Survey," the data source should be recorded as "2018 Current Population Survey." If the first reference to this data source is found on page 1 of the appendix, you should record its location as "A1." Do this for all data sources mentioned in the paper. Each row represents a unique data source.

Data sources also vary by unit of analysis, with some sources matching the same unit of analysis used in the paper (as in previous examples). In contrast, others may be less clear, e.g., "our information on regional minimum wages comes from the Bureau of Labor Statistics." You should record such data source as "regional minimum wages from the Bureau of Labor Statistics."

Next, look at the reproduction package and map the data sources mentioned in the paper to the data files in the reproduction package. In the Location column, record their folder locations relative to the main reproduction folder. In addition to looking at the existing data files, we recommend that you also review the first lines of all code files (especially cleaning code), looking for lines that call the datasets. Inspecting these scripts may help you understand how different data sources are used, and possibly identify any missing files from the reproduction package. Whenever a data source contains multiple files, enter them in the same cell, separated by semicolon (;).

If you cannot find the files names corresponding to a specific data source, type "Not available" in the Data Files field. If you can identify the file name. Check the Provided column if the data source was included in the original reproduction package. Check the Cited column if the data source was explicitly cited in the paper. Record your work using the following structure:

Universidad de los Andes
Facultad de Economía

ECONOMIA URBANA

2022-1

Econ 4683

Ignacio Sarmiento-Barbieri

Data.Source	Page	Data.Files	Location	Provided	Cited
"Current Population Survey 2018"	A1	cepr_march_2018.dta	data/	TRUE	FALSE
"Provincial Administration Reports"	A4	coast_simplepoint2.csv; rivers_simplepoint2.csv; RAIL_dummies.dta;	Data/maps/	TRUE	FALSE
"2017 SAT scores"	4	Not available	data/to_clean/	FALSE	TRUE

Note on data citations: In your assessment of whether or not a data source has been cited should take into account the general guidelines for what is considered a complete data citation. In-text citations, mention of the source location (e.g. url to source location), and description of the source or the data alone are examples of incomplete data citations. Please refer to the Guidance on Data Citations by Social Science Data Editors for a comprehensive overview and more reference material on data citation standards for the social sciences.

1.2 Description of the analytic data

For this part, you are going to describe the analytic data. For this part you are going to construct a table that looks like this:

Analytic.Data	Location	Description
final_data.csv	analysis/fig1/	data for figure1
all_waves.csv	final_data/v1_april/	data for region-level analysis

To fill this table follow these steps:

- For the first column, the Analytic Data column, identify all analytic data files in the reproduction package and record their names. You will recognize analytic data files based on the documentation, their location folder, or if a code file either uses them to compute a statistic that is displayed in the paper (or appendix).
- Second, in the Location column, record each analytic data file's location relative to the main folder of the reproduction package .
- Finally, provide a one-line description of each file in the Description column (e.g., all_waves.csv can be "data for region-level analysis").

Universidad de los Andes
Facultad de Economía

ECONOMIA URBANA

2022-1

Econ 4683

Ignacio Sarmiento-Barbieri

1.3 Description of the code scripts

The last part of this section consists in describing the code scripts. As with the other two subsections the output is a table, that looks like this:

file_name	location	inputs	outputs	description	primary_type
output_table1.do	code/analysis/	analysis_data01.csv	output1_part1.txt	produces first part of table 1 (unformatted)	analysis
data_cleaning02.R	code/cleaning	admin_01raw.csv	analysis_data02.csv	removes outliers and missing vals from raw admin data	cleaning

To fill this table follow these steps:

- First, identify all code files in the reproduction package and record their names in the File Name column and record their locations relative to the main folder in the Location column.
- Then, review the beginning and end of each code file to identify the inputs required to successfully run the file and the outputs it produces. Inputs are data sets or other code scripts that are typically found at the beginning of the script (e.g., load, read, source, run, do). Outputs are other data sets, or plain text files typically found at the end of a script (e.g., save, write, export). Record those in the Inputs and Outputs columns.
- Finally, provide a brief description of the code's function in the Description column and classify its function as analysis or cleaning and/or construction in the Primary Type column.

Note: If there are many files for a given data source, analysis data or code scripts, that have a recursive structure (e.g., wages0001.csv, ... , wages0734.csv), record only one generic entry that represents all these of files (e.g., wages0XYZ.csv).

2 Assign a reproducibility score.

Once you have identified all possible inputs and have a clear understanding of the connection between the display items and their inputs, you can assign reproducibility scores to individual display items. *Please choose one of the following levels and provide a brief justification for your score*

Universidad de los Andes
Facultad de Economía

ECONOMIA URBANA

2022-1

Econ 4683

Ignacio Sarmiento-Barbieri

Table 1: Reproducibility Scores

Level 1 (L1)	No data or code are available. Possible improvements include adding: raw data, analysis data, cleaning code, and analysis code.
Level 2 (L2)	Code scripts are available (partial or complete), but no data are available. Possible improvements include adding: raw data and analysis data.
At this stage the paper you chose should be above the first two levels	
Level 3 (L3)	Analytic data and code are partially available, but raw data and cleaning code are missing. Possible improvements include: completing analysis data and/or code, adding raw data, and adding analysis code.
Level 4 (L4)	All analytic data sets and analysis code are available, but the code fails to run or produces results inconsistent with the paper. Possible improvements include: debugging the analysis code or obtaining raw data.
Level 5 (L5)	Analytic data sets and analysis code are available and they produce the same results as presented in the paper. The reproducibility package may be improved by obtaining the original raw data.
Level 6 (L6)	Cleaning code scripts are available (partial or complete), but raw data is missing. Possible improvements include: adding raw data.
Level 7 (L7)	Cleaning code is available and complete, and raw data is partially available. Possible improvements: adding raw data.
Level 8 (L8)	All the materials (raw data, analytic data, cleaning code, and analysis code) are available. However, the cleaning code fails to run or produces different results from those presented in the paper or the analysis code fails to run or produces results inconsistent with the paper . Possible improvements: debugging the cleaning or analysis code.
Level 9 (L9)	All the materials (raw data, analytic data, cleaning code, and analysis code) are available. The analysis code produces the same output as presented in the paper. However, the cleaning code fails to run or produces different results from those presented in the paper. Possible improvements: debugging the cleaning code.
Level 10 (L10)	All necessary materials are available and produce consistent results with those presented in the paper. The reproduction involves minimal effort and can be conducted starting from the analytic data (CRA) and the raw data (CRR). Note that Level 10 is aspirational and may be unattainable for most research published today.

3 Bonus

Connect display items to all its inputs using the Diagram Builder in the SSRP.