# Structure of the Reproduction Activity

(1) (2) (3) (4) (5)  
 scope --> assess --> improve --> robust --> extend  
 ▲ |  
 | |  
 |\_\_\_\_\_\_\_\_\_|  
  
 Suggested level  
 of effort:  
 - Undergraduate  
 thesis: 10% 30% 40% 20% 0%  
 - Graduate  
 course: 10% 25% 20% 40% 5%  
 - Graduate  
 research: 5% 10% 5% 10% 70%

1 - Read the paper and identify main scientific claims. Define scope of the reproduction exercise (10%). Declare if assessment will be about [all or main outputs](01_definitions.md#types-of-reproduction).

2 - Assess the current level of computational reproducibility of the different outputs in the paper. - 2.1 - List all the input materials found or referred in the reproduction package:  
- 2.1.1 - List all raw data sources as described in the paper (for example “CPS March 2018 from CEPR”) and, if raw data is available, match it with the filename in the reproducibility package (cepr\_march\_2018.dta). The result will have the following structure:

Raw data information:  
 |----------------------|------|-----------------------------------------------|----------|---------------------|  
 | data\_source | page | data\_files | complete | known\_missing |  
 |----------------------|------|-----------------------------------------------|----------|---------------------|  
 | "Current Population | 3 | cepr\_march\_2018.dta | yes | |  
 | Survey 2018" | | | | |  
 |----------------------|------|-----------------------------------------------|----------|---------------------|  
 | "DHS 2010 - 2013" | 4 | nicaraguaDHS\_2010.csv; | no | boliviaDHS\_2011.csv |  
 | | | boliviaDHS\_2010.csv; nicaraguaDHS\_2011.csv; | | |  
 | | | nicaraguaDHS\_2012.csv; boliviaDHS\_2012.csv; | | |  
 | | | nicaraguaDHS\_2013.csv; boliviaDHS\_2013.csv | | |  
 |----------------------|------|-----------------------------------------------|----------|---------------------|  
 | "2017 SAT scores" | 4 | Not available | no | |  
 |----------------------|------|-----------------------------------------------|----------|---------------------|  
 | ... | ... | ... | ... | ... |  
 |----------------------|------|-----------------------------------------------|----------|---------------------|  
  
- 2.1.2 - List and describe all the analytic data sets. The resulting report will have the following structure:   
  
 Analysis data information:  
 |----------------|-----------------------|--------------------------------|  
 | analysis\_data | location | description |  
 |----------------|-----------------------|--------------------------------|  
 | final\_data.csv | /analysis/fig1/ | data for figure1 |  
 |----------------|-----------------------|--------------------------------|  
 | all\_waves.csv | /final\_data/v1\_april/ | data for region-level analysis |  
 |----------------|-----------------------|--------------------------------|  
 | ... | ... | ... |  
 |----------------|-----------------------|--------------------------------|  
  
- 2.1.3 - List and describe all the code scripts. The result will have the following structure.   
  
 |-------------------|------------------|---------------------|---------------------|----------------------|--------------|  
 | file\_name | location | inputs | outputs | description | primary\_type |  
 |-------------------|------------------|---------------------|---------------------|----------------------|--------------|  
 | output\_table1.do | /code/analysis/ | analysis\_data01.csv | output1\_part1.txt | produces first part | analysis |  
 | | | | | of table 1 | |  
 | | | | | (unformatted) | |  
 |-------------------|------------------|---------------------|---------------------|----------------------|--------------|  
 | data\_cleaning02.R | /code/cleaninig/ | admin\_01raw.csv | analysis\_data02.csv | removes outliars | cleaning |  
 | | | | | and missing vals | |  
 | | | | | from raw admin data | |  
 |-------------------|------------------|---------------------|---------------------|----------------------|--------------|  
 | ... | ... | ... | ... | ... | ... |  
 |-------------------|------------------|---------------------|---------------------|----------------------|--------------|

* 2.2 - Draw diagrams from output to raw data sources. For more examples of diagrams connecting final output to initial raw data, [see here](CREATE%20FILE).
* table 1  
   └───[code] formatting\_table1.R  
   ├───output1\_part1.txt   
   | └───[code] output\_table1.do   
   | └───[data] analysis\_data01.csv  
   | └───[code] data\_cleaning01.R  
   | └───[data] survey\_01raw.csv  
   └───output1\_part2.txt   
   └───[code] output\_table2.do   
   └───[data] analysis\_data02.csv  
   └───[code] data\_cleaning02.R  
   └───[data] admin\_01raw.csv
* This diagram can be represented in data format by specifying how each component depends to its inputs. For example:
* Data representation of diagram behind Table 1.  
   |-------|-------------------|---------------------|------------|  
   | order | component | depends\_on | other\_info |  
   |-------|-------------------|---------------------|------------|  
   | 1 | table1 | formatting\_table1.R | ... |  
   |-------|-------------------|---------------------|------------|  
   | 2 | table1 | output1\_part2.txt | ... |  
   |-------|-------------------|---------------------|------------|  
   | 3 | output1\_part1.txt | output\_table1.do | ... |  
   |-------|-------------------|---------------------|------------|  
   | 4 | output\_table1.do | analysis\_data01.csv | ... |  
   |-------|-------------------|---------------------|------------|  
   | ... | ... | ... | ... |  
   |-------|-------------------|---------------------|------------|
* 2.3 - Assess the current [level](README.md#levels-of-computational-reproducibility) of computational reproducibility of each output
* |-------------|-------|-----------------------|------------|  
   | output\_name | level | desc\_of\_missing\_files | other\_info |  
   |-------------|-------|-----------------------|------------|  
   | table 1 | 4 | | ... |  
   |-------------|-------|-----------------------|------------|  
   | table 2 | 7 | | ... |  
   |-------------|-------|-----------------------|------------|  
   | figure 1 | 5 | | ... |  
   |-------------|-------|-----------------------|------------|  
   | ... | ... | ... | ... |  
   |-------------|-------|-----------------------|------------|
* 2.4 - Check for the existence of paper-level best practices for reproducibility
* |--------------------------|------------|  
  | best\_practice | exist |  
  |--------------------------|------------|  
  | version control | yes |  
  | dynamic document | no |  
  | translate to open source | no |  
  | file organization | yes |  
  | computing capsule | no |  
  |--------------------------|------------|

3 - Improve the current computational reproducibility of [all | main | specific] output of the paper. Choose among the feasible and desired types of improvements described in this [section](01_definitions.md#possible-improvements)  
- 3.1 - Level-specific quality improvements: add data/code, debug code.

| output\_name | imprv | name\_description\_of\_added\_files | lvl |  
 |-------------|-------|-----------------------------------|-----|  
 | table 1 | +AD | ADD EXAMPLES | 5 |  
 | table 1 | +RD | ADD EXAMPLES | 5 |  
 | table 1 | DCC | ADD EXAMPLES | 5 |  
 | figure 1 | +CC | | 6 |  
 | figure 1 | DAC | | 6 |  
 | inline 1 | DAC | | 8 |  
 | ... | ... | ... | ... |

* 3.2 - Overall quality improvement: folders, literate programming, open source statistical software, version controls software, computing capsule.
* | improvement | took\_place | description |  
   |--------------------------|------------|----------------------|  
   | version control | yes | ADD EXAMPLES |  
   | dynamic document | no | - |  
   | translate to open source | no | - |  
   | file organization | yes | ADD EXAMPLES |  
   | computing capsule | yes | ADD EXAMPLES |

4 - Check the robustness of results to variations in the underlying set of analytical choices  
- 4.1 - Identify all possible analytical choices.

| file\_name | line\_number | choice\_type | choice\_value | Source |  
 |------------|-------------|---------------------|--------------------------------|---------------------|  
 | code\_01.do | 73 | data subsetting | males | original |  
 | code\_01.do | 122 | variable definition | income = wages + capital gains | "code\_01.do-L103" |  
 | code\_05.R | 143 | controls | age, income, education | original |  
 | ... | ... | ... | ... | ... |

* 4.2 Test the robustness of results to alternative (sensible) specifications
  + Identify sensible alternatives to analytical choices.
  + Sample from sensible analytical choices and re-run: report how much do results change as fraction of standard deviations.
  + Jackknife the preferred estimate.
  + Use ML to select among covariates…

5 - [Optional] Propose extensions that can be carried out with these data/methods.

# Final products

* One-page introduction describing why you chose this paper
* Two-page summary of paper
* 2 Completed surveys:
  + i - General information about the paper and specific information about output to reproduce.
  + ii - Assessment of how (computationally) reproducible is the paper; description of improvements to its reproducibility; record of all the analytical choices identified in the exercise.
* ACRE report card with all the improvements that were created by the researcher reproducing the paper. The list of improvements will be made public and original authors will receive a copy of the report card. The option of anonymity will be provided to the researchers reproducing the paper.
* New Readme file (autogenerated).
* Data with all analytical choices identified.
* ?? Narrated description of improvements to original CR of the paper, assessment of robustness of results. Lessons from the exercise and possible future extensions.