# Le rôle des tiers de confiance dans la vérification de la reproductibilité : L'exemple de cascad
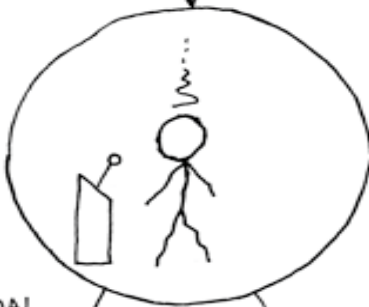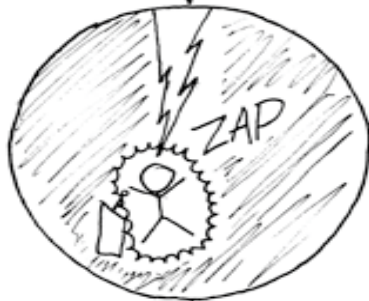
Christophe Pérignon

Professor of Finance and Associate Dean
for Research at HEC Paris

Founder/Head of cascad (UAR 2007)

January 31, 2025

Réseau Français de la
Recherche Reproductible

✓cascad

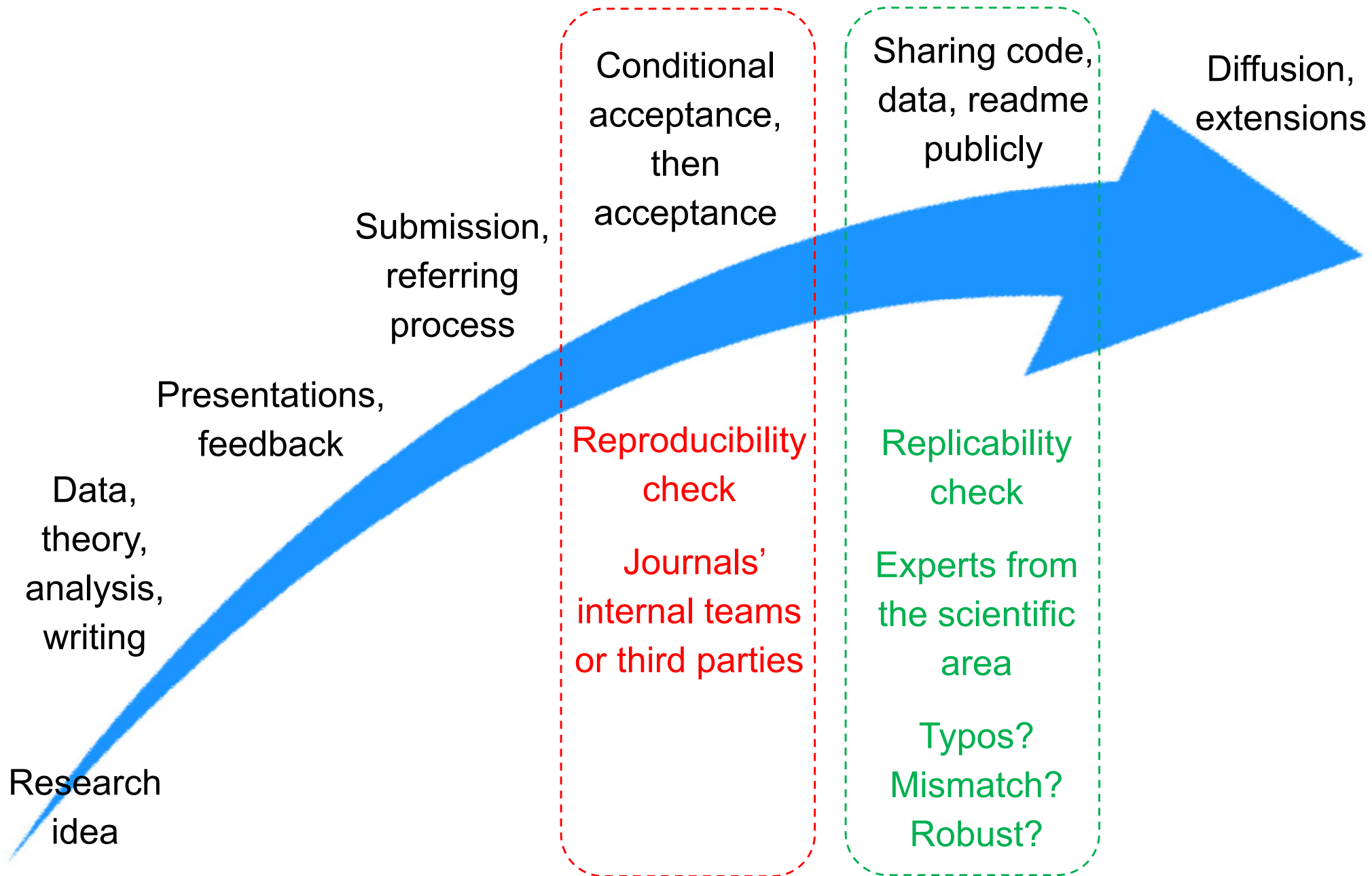**Reproducibility** or **Computational Reproducibility** is defined as:

> same data + same method = same results

**Replicability** is defined as:

> **new** data + same method = same results
>
> same data + **new** method = same results
>
> **new data** + **new** method = same results

Barba (2020)

✔cascad

**Scientific Process**

Research idea

Data, theory, analysis, writing

Presentations, feedback

Submission, referring process

Conditional acceptance, then acceptance

Reproducibility check

Journals' internal teams or third parties

Sharing code, data, readme publicly

Replicability check

Experts from the scientific area

Typos? Mismatch? Robust?

Diffusion, extensions

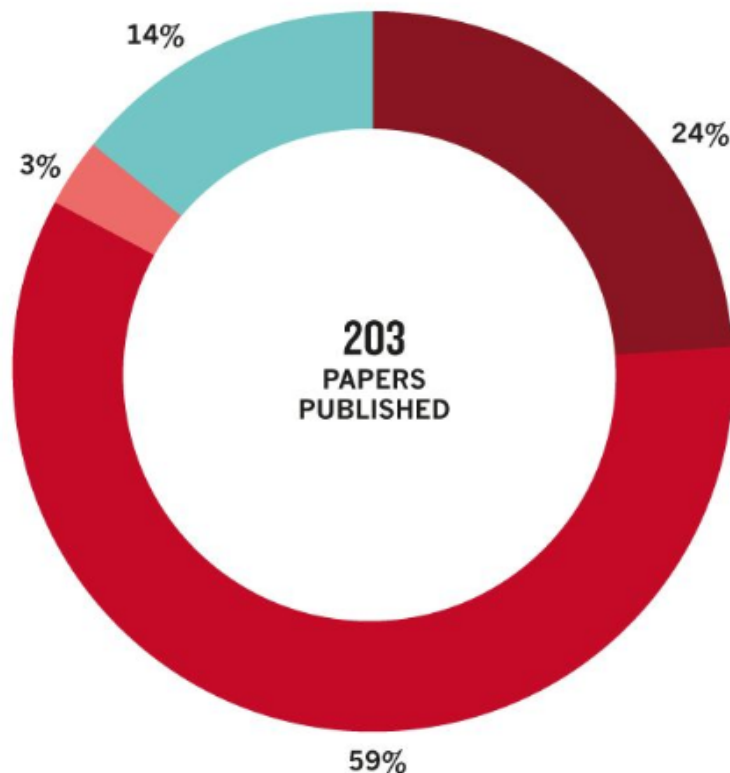# ✓cascad    Reproducibility rate in economics is low

## Gertler et al. (Nature, 2018)

### REPLICATION RARELY POSSIBLE
An analysis of 203 economics papers found that fewer than one in seven supplied the materials needed for replication.
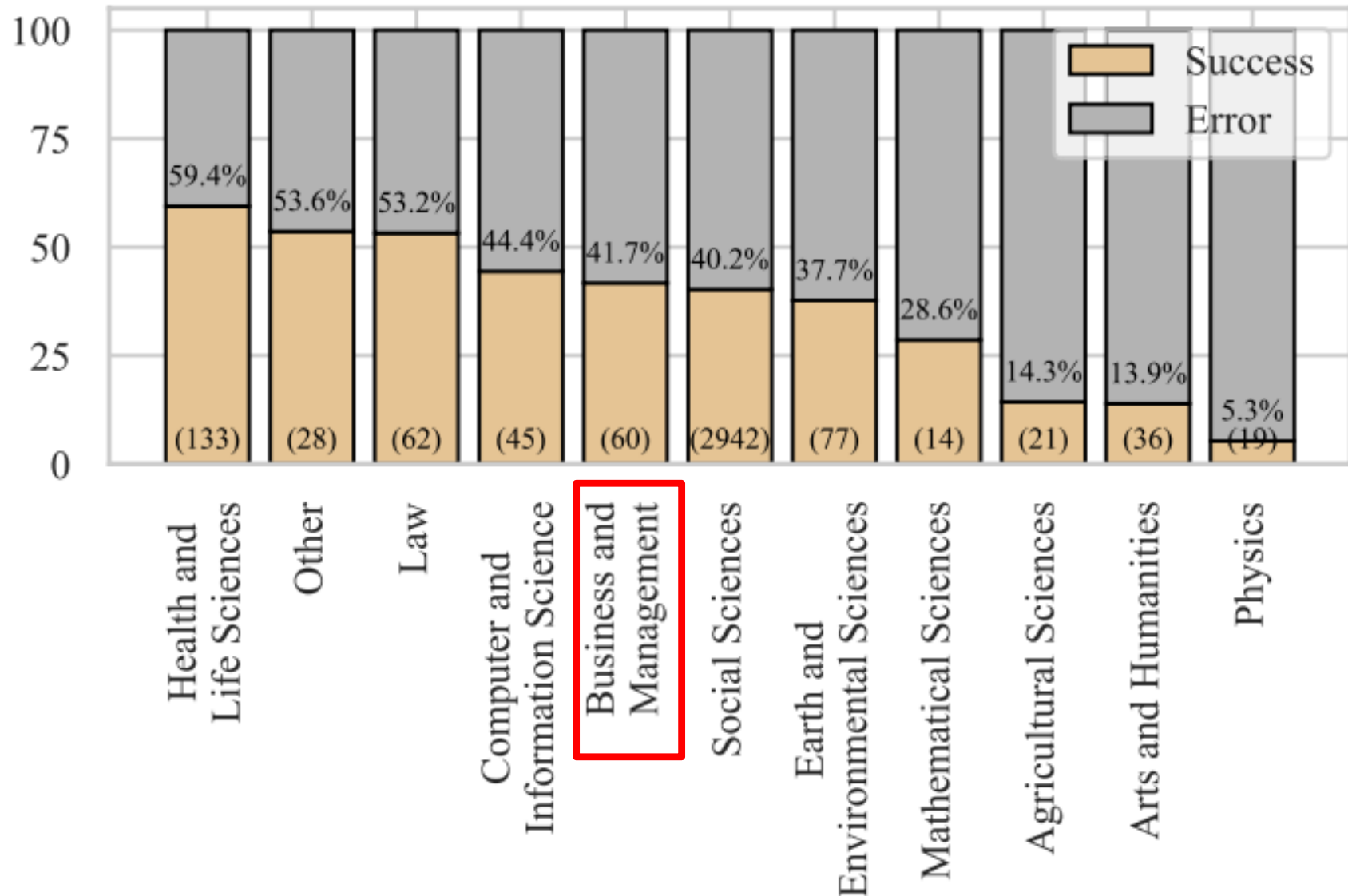
**ELEMENTS PROVIDED*:**
- ■ None
- ■ One or more missing
- ■ All, code doesn't run
- ■ All, code runs

14%

3%

24%

**203 PAPERS PUBLISHED**

59%

## Chang and Li (AER, 2017)

Only able to reproduce the results for one-third of 67 papers published in top economic journals from the code and data available on the journals' repositories.

Another 10% of the papers were reproduced using the help of the authors

# Benchmarking with other disciplines



Re-execution rate per field of study

Source: Trisovic, Lau, Pasquier and Crosas (*Nature Scientific Data*, 2022)

- First crowd-sourcing study in finance: the Finance Crowd Analysis Project, or in short #fincap (Menkveld et al., 2024)

- 168 research teams from 37 countries were instructed:
  - to analyze the same dataset of EuroStoxx 50 futures transactions (720 mio transactions)
  - to answer the same six research questions

- Pérignon et al. (2024) use the teams' computer code and the common raw dataset to try to reproduce the results provided by all teams to all research questions:

$$168 \times 6 = 1,008 \text{ empirical findings}$$

# ✓cascad  Step 1. Level of computational reproducibility

- One full-time, experienced reproducibility verificator supervised by two reproducibility editors (cascad)

- Virtually all software, unlimited CPU, can seek help from colleagues, but one week max and no contact with authors

- For each question, two coefficients: beta + standard-error

| Reproducibility Rating | Reproducibility Score | Difference |
|:---:|:---:|:---:|
| RRR | 100 | None, perfect accuracy |
| RR | 75 | Only small differences |
| R | 50 | One large difference |
| D | 25 | Several large differences |
| DD | 0 | No result generated |

✓cascad

- Result level:

✔cascad
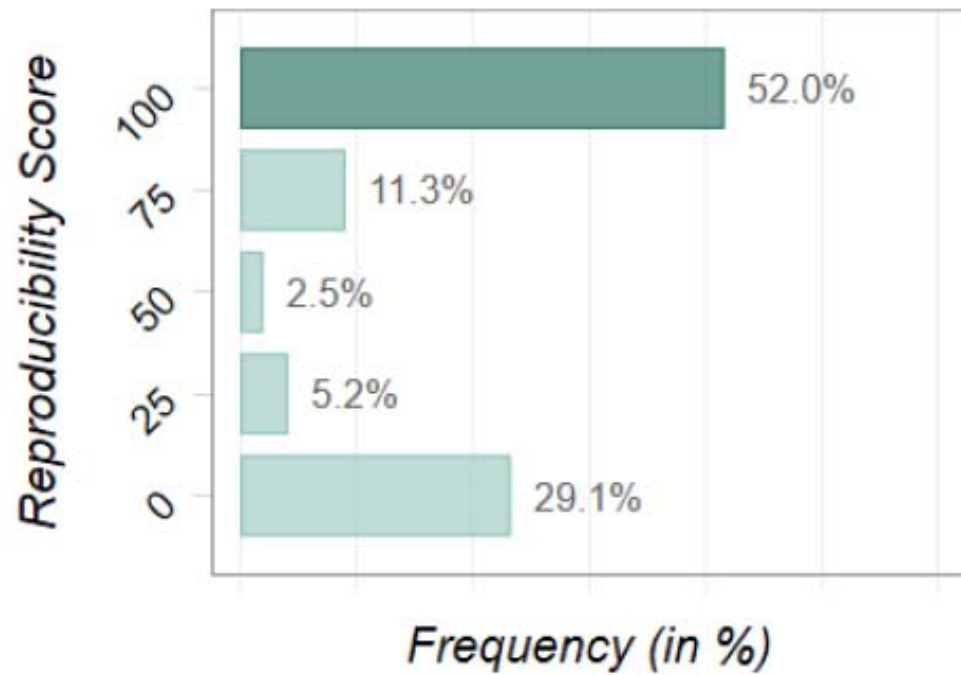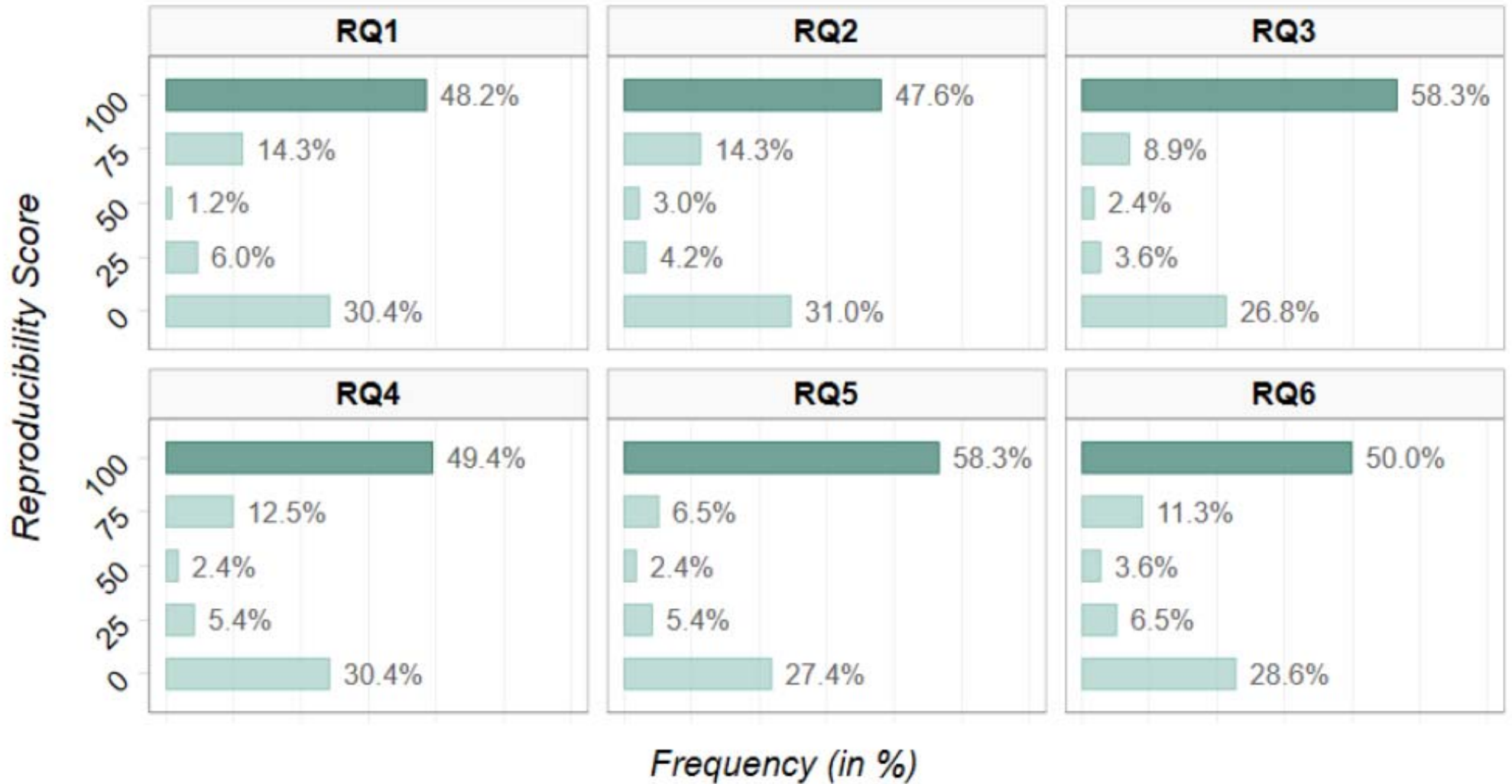
- Result level:
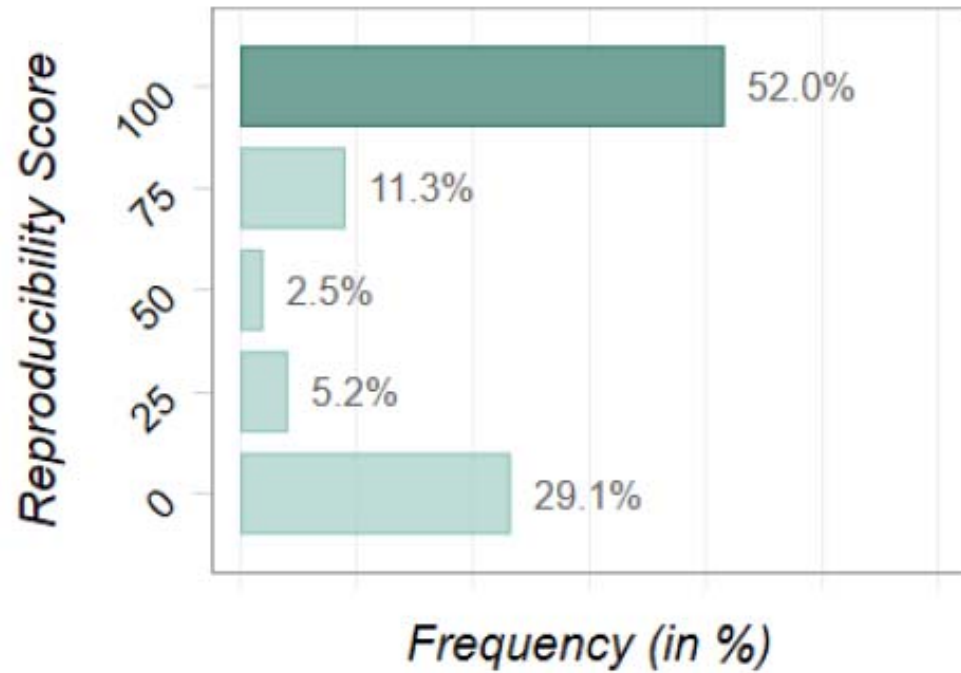
- Team level:
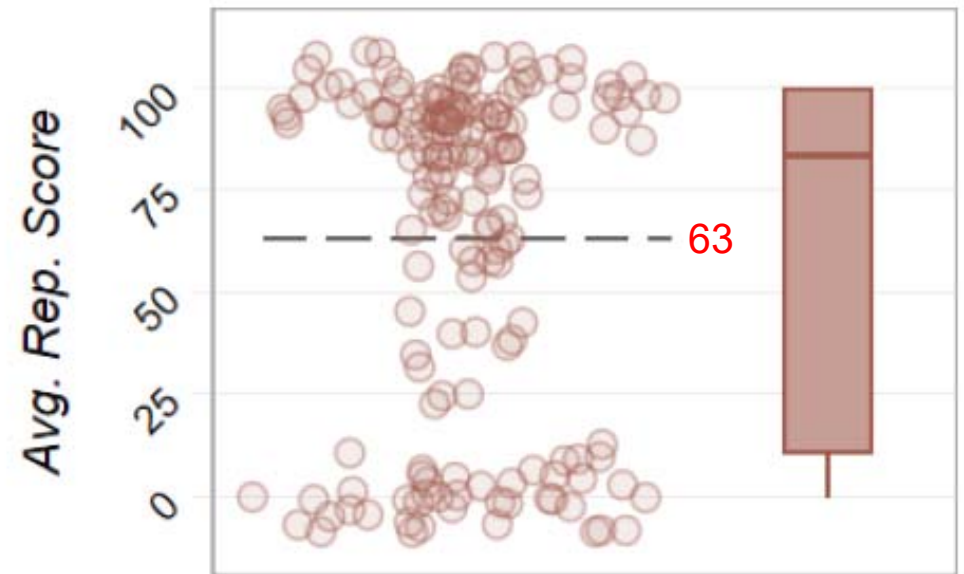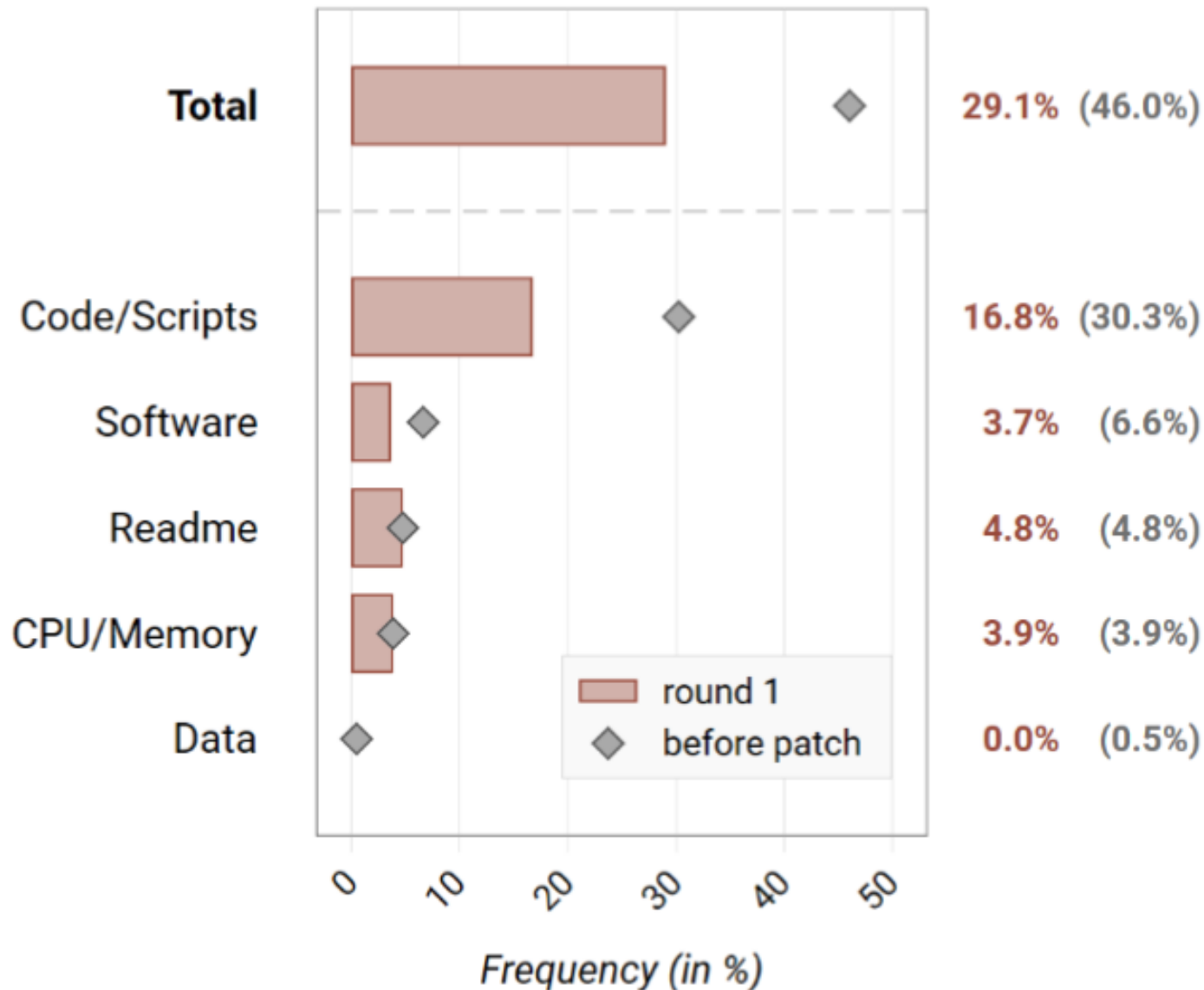
| Type | Problems | Examples | #fincap | |
| --- | --- | --- | --- | --- |
| | | | *R0* | *R1* |
| CPU/Memory | Time constraint exceeded | Code fails to complete within 168 hours (1 week) | 2.5 | 2.5 |
| | Insufficient CPU/memory | Code requires more than 256 GB of RAM | 1.4 | 1.4 |
| Code/Scripts | Improperly named variables | Code uses one variable that do not exist or one which is misspelled | 10.8 | 5.7 |
| | Code fails to import the raw data | Python code attempts to load the data but cannot parse it | 5.8 | 4.2 |
| | Data conversion failure | Code fails to convert the .csv into .Rdata files | 3.0 | 3.0 |
| | Missing/unreadable scripts | Scripts become unreadable due to conversion errors from MAC | 2.1 | 2.1 |
| | Code runs but does not produce results | SAS code runs but does not generate results for one #*fincap* hypothesis | 1.6 | 1.6 |
| | Problems with intermediary results | Intermediary results generated with SAS but R fails to use them | 3.3 | 0.3 |
| | Lines of code must be added/removed | Part of the code was bugged but not used to generate results | 3.8 | — |
| | Code fails to compile | Fortran code does not generate any executable file | — | — |
| Data | Missing data | Missing required dataset from the American Time Use Survey (ATUS) | — | — |
| | Altered data | Data provider (French Customs) updated the raw data | — | — |
| | Restricted data access | No access to National Center for Education Stat. (NCES) outside the US | — | — |
| | Inappropriate data format | Code expected .dta files instead of .csv | 0.5 | — |
| Readme | Information to map output/results | Missing explanation on where to find results in a 100-page log file | 3.0 | 3.0 |
| | Information about code/software | Missing command to run a given Fortran code | 1.8 | 1.8 |
| | Information about data access | Missing instructions to download specific data from IPUMS platform | — | — |
| Software | Incompatible environment | Inability to run SAS code in Linux | 1.8 | 1.8 |
| | Unavailable libraries/software | Specific R library removed from the CRAN repository | 1.9 | 1.9 |
| | Versioning | Code runs in Matlab R2019b but not in more recent versions | 3.0 | — |
| **Total** | | | **46.1** | **29.1** |

We consider six dimensions of scholarship:

(i) researchers' characteristics and skills

(ii) type of research question addressed

(iii) software used

(iv) complexity of the computer code

(v) quality of the research

(vi) effort exhibited by researchers

| Logit regressions | Model (1) | | Model (2) | |
| --- | --- | --- | --- | --- |
| *Academic Quality:* | | | | |
| » PC-1 | −0.021 | (0.319) | | |
| » Seniority[†] | | | −0.152 | (0.070) |
| » Top Publication[†] | | | 0.090 | (0.287) |
| » Citations (in logs) | | | −0.004 | (0.853) |
| » Expertise (0–10) | | | 0.009 | (0.691) |
| *Coding Skills:* | | | | |
| » PC-1 | 0.092* | (0.001) | | |
| » Parallel Comp.[†] | | | 0.107 | (0.319) |
| » Loops/Matrix Operations[†] | | | 0.330* | (0.000) |
| » Large Data[†] | | | 0.001 | (0.982) |
| » Coding Skills[†] | | | −0.015 | (0.832) |
| *Coauthor:* | | | | |
| » Team of Two[†] | 0.052 | (0.524) | 0.038 | (0.634) |
| *Gender:* | | | | |
| » Female[†] | −0.059 | (0.408) | −0.046 | (0.523) |
| *Location:* | | | | |
| » North America[†] | −0.010 | (0.901) | −0.024 | (0.765) |
| » Asia-Pacific[†] | −0.138 | (0.168) | −0.161 | (0.104) |
| » Other countries[†] | 0.036 | (0.755) | 0.085 | (0.478) |
| *Software:* | | | | |
| » Matlab[†] | −0.080 | (0.473) | −0.069 | (0.507) |
| » Python[†] | 0.014 | (0.884) | 0.014 | (0.881) |
| » R[†] | −0.052 | (0.552) | −0.063 | (0.464) |
| » SAS[†] | 0.130 | (0.131) | 0.147 | (0.095) |
| » Stata[†] | −0.116 | (0.132) | −0.110 | (0.146) |
| *Research Questions:* | | | | |
| » RQ2[†] | −0.006 | (0.862) | −0.006 | (0.862) |
| » RQ3[†] | 0.101* | (0.003) | 0.101* | (0.003) |
| » RQ4[†] | 0.012 | (0.732) | 0.012 | (0.732) |
| » RQ5[†] | 0.101* | (0.006) | 0.101* | (0.005) |
| » RQ6[†] | 0.018 | (0.613) | 0.018 | (0.613) |

|  | Model (3) | | Model (4) | |
| --- | --- | --- | --- | --- |
| *Research Quality:* | | | | |
| » Peer Evaluation (0–10) | 0.019 | (0.197) | 0.017 | (0.248) |
| » Out-of-Consensus Result[†] | −0.235* | (0.010) | −0.209* | (0.014) |
| *Code Complexity:* | | | | |
| » PC-1 | −0.045* | (0.048) | | |
| » Number of Software | | | −0.083 | (0.209) |
| » Number of Script Files | | | 0.000 | (0.975) |
| » Size of Software (in kb) | | | 0.000 | (0.188) |
| » Actual CPU Time (in minutes) | | | 0.000 | (0.520) |
| » Lack of Master File[†] | | | 0.029 | (0.686) |
| » Help from Verifier[†] | | | 0.065 | (0.414) |
| *Documentation Quality:* | | | | |
| » PC-1 | 0.046* | (0.041) | | |
| » Readme File[†] | | | 0.090 | (0.830) |
| » Size of Readme File (in kb) | | | −0.005 | (0.743) |
| » Software Requirements[†] | | | −0.047 | (0.846) |
| » Runtime[†] | | | 0.046 | (0.675) |
| » Computer Specification[†] | | | 0.114 | (0.439) |
| » Instructions to Verifier[†] | | | 0.211 | (0.478) |
| » Mapping Output/Results[†] | | | 0.125 | (0.151) |

# Step 4. Expected level of reproducibility

"Do you think it would be possible to reproduce your results from the raw data and your computer code?"

A. "One would find exactly the same results"
B. "Only minor differences may arise"
C. "Major differences may arise"
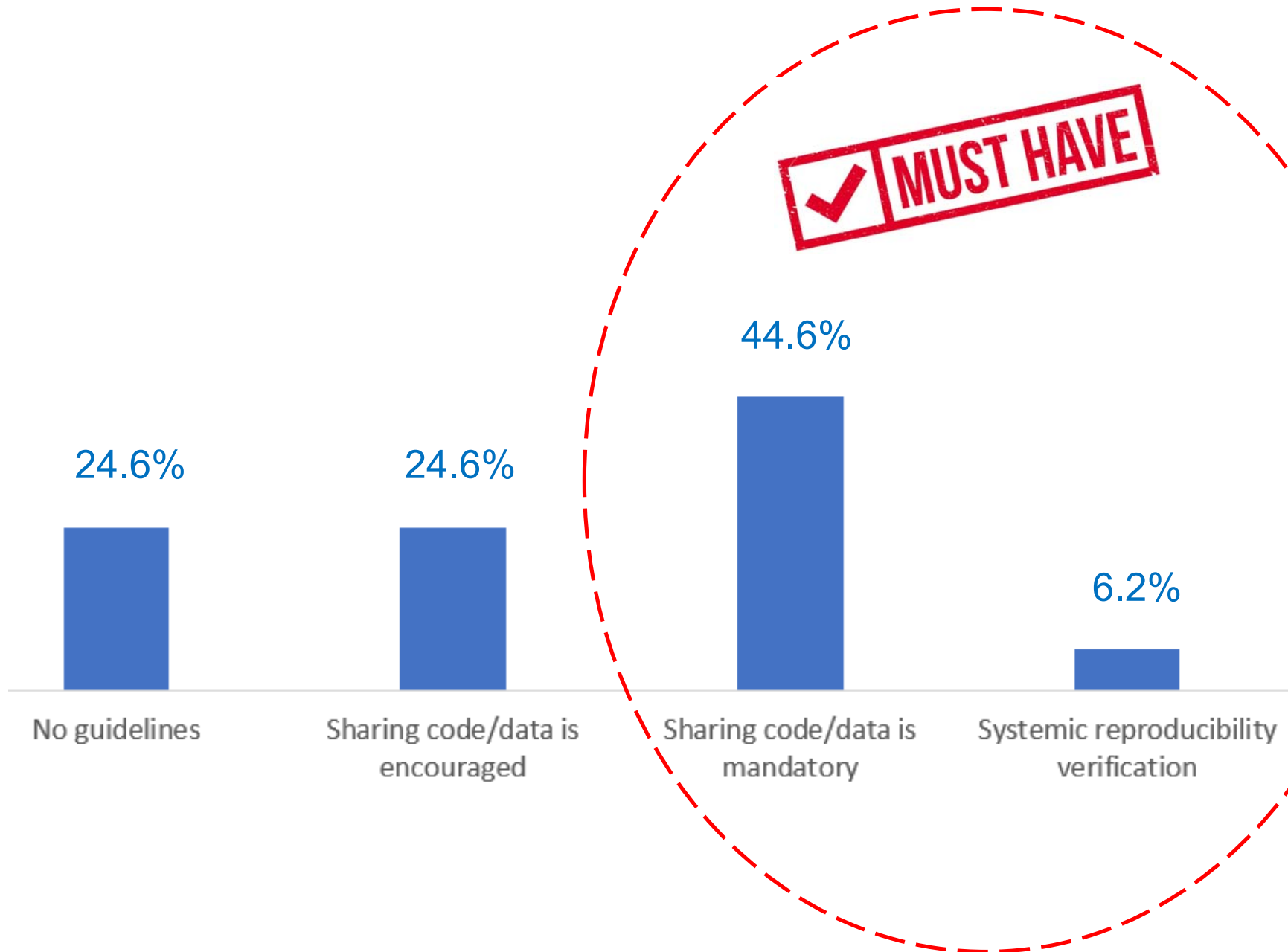D. "It would be impossible to reproduce the results"
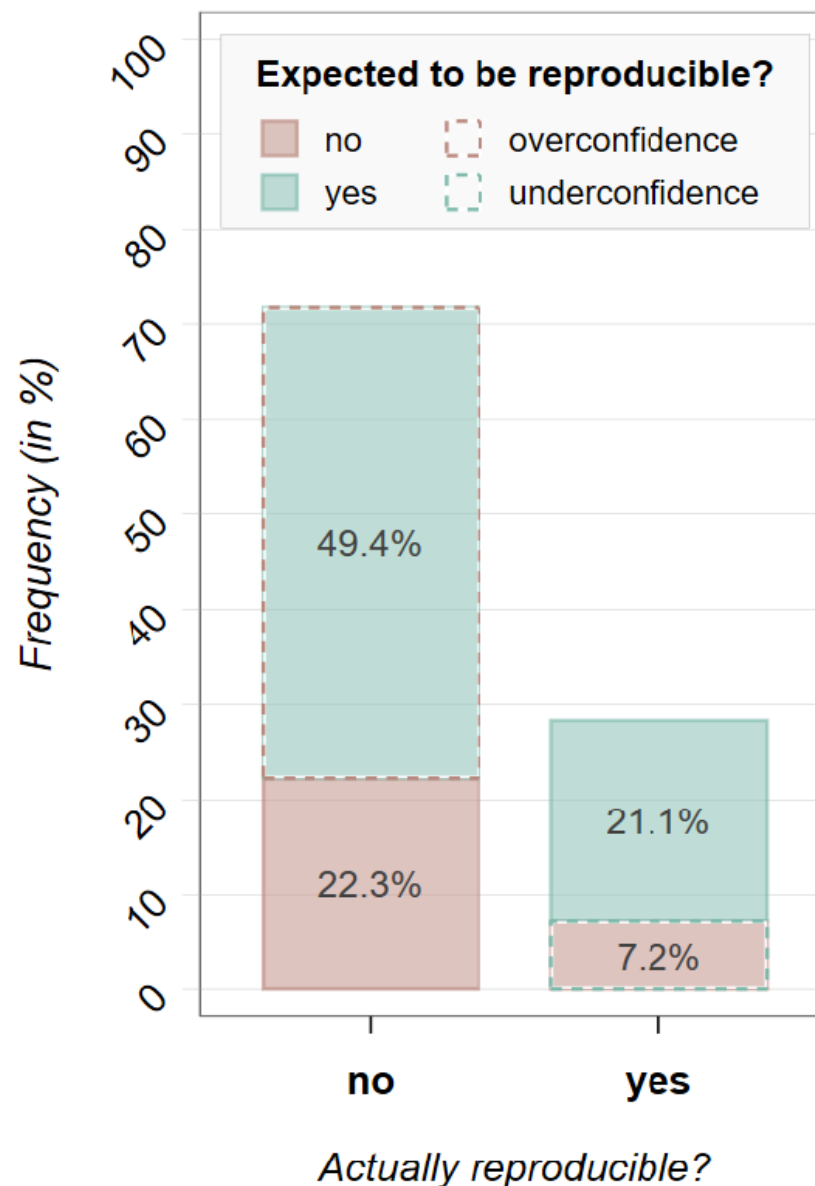
# Step 5. Expected level of difficulty

"How easy would it be to reproduce your results?"

A. "Straightforward"
B. "Quite easy"
C. "Challenging"
D. "Very difficult"
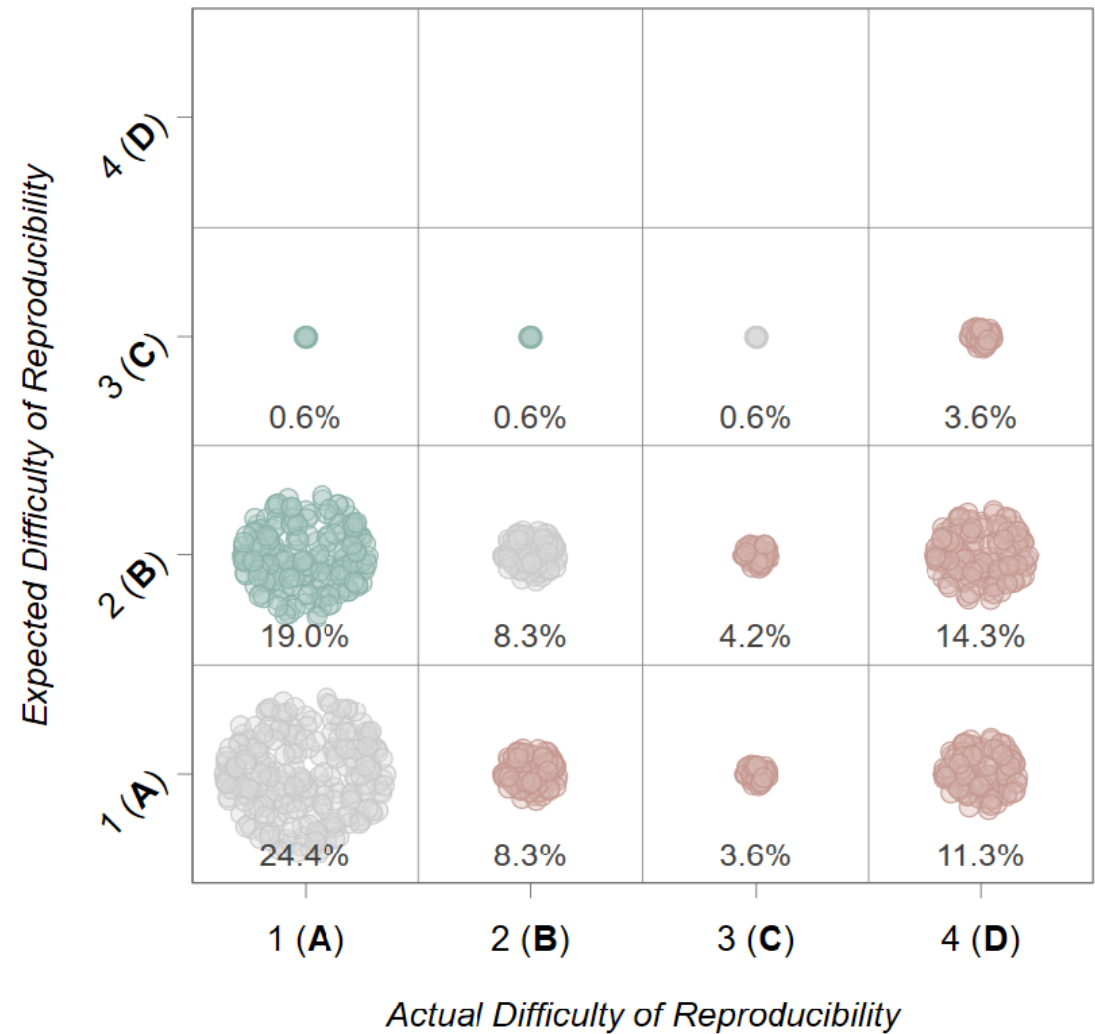
# Best academic journals

American Economic Review (Econ #1)　　cascad

Journal of Political Economy (Econ #2)　　cascad

Quarterly Journal of Economics (Econ #3)

Review of Economics Studies (Econ #4)

Econometrica (Econ #5)　　cascad

Management Science (QuantMgt #1)　　cascad

American Journal of Political Science (PoSc #1)　　cascad

Journal of the American Statistical Association (Stat #1)

Systemic Verification: 

Collaboration with cascad: cascad

# Third-party reproducibility verificator

![cascad logo]

## cascad = **C**ertification **A**gency for **S**cientific **C**ode and **D**ata

The cascad reproducibility certification attests that the numerical results reported in a scientific article can be reproduced from a set of numerical resources (code and data) provided by the authors.
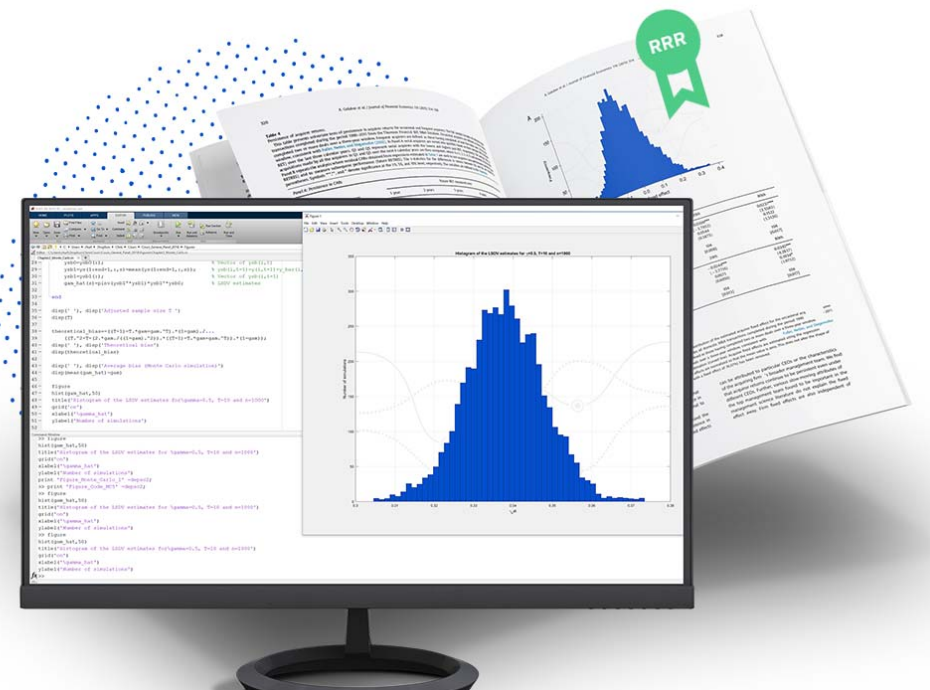


- CNRS, HEC Paris, U. Orléans
- www.cascad.tech
- Founded by researchers for researchers (and journals)
- 400 verifications

Link:

https://www.youtube.com/watch?v=i17UI2bKh0E&feature=youtu.be

cascad for researchers

Author requests a certification

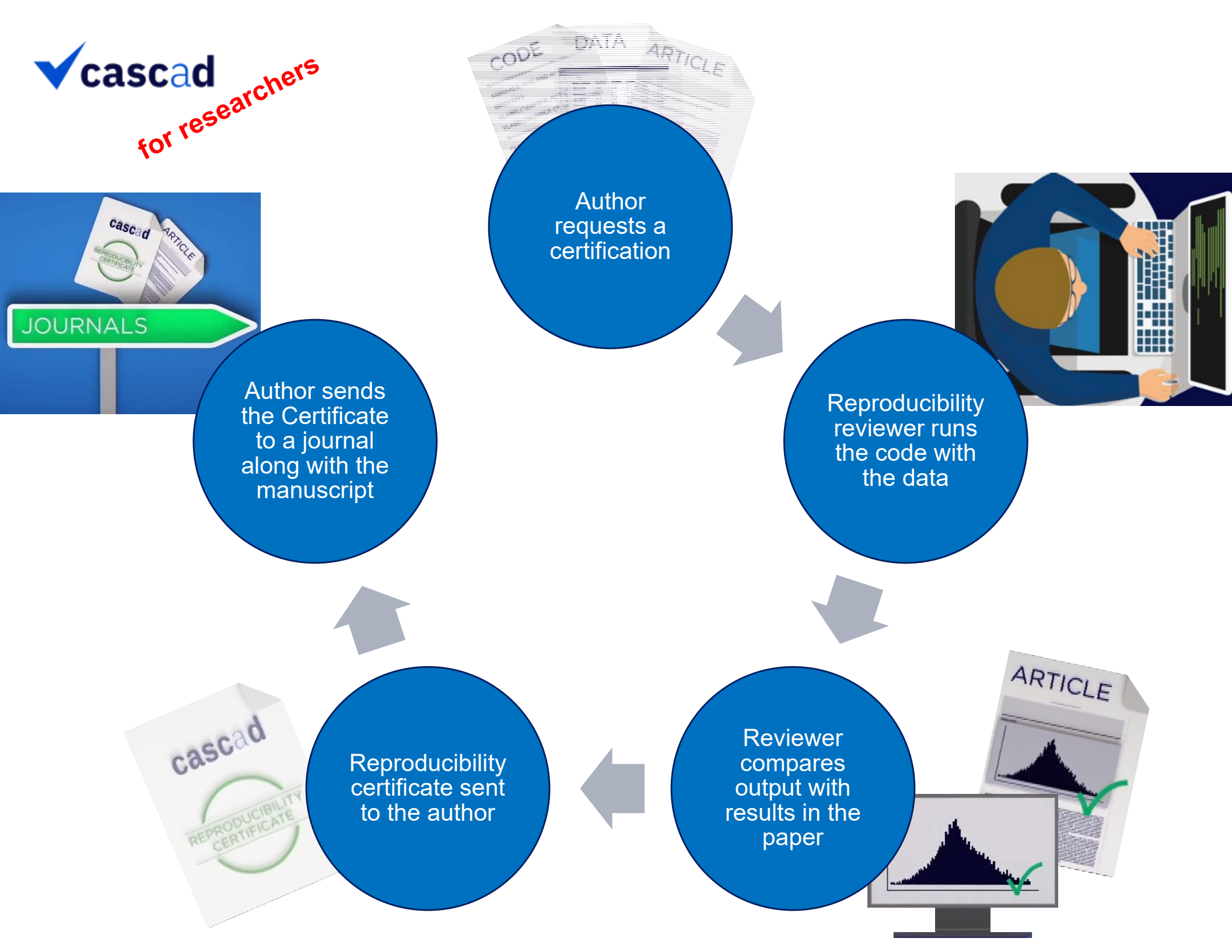Reproducibility reviewer runs the code with the data

Reviewer compares output with results in the paper

Reproducibility certificate sent to the author

Author sends the Certificate to a journal along with the manuscript

cascad for journals

Journal requests a pre-publication verification

Reproducibility reviewer runs the code with the data

Reviewer compares output with results in the paper

Reproducibility report sent to the data editor

Data editor makes the final decision: publish or ask for more info/code/data

![cascad logo] **✓cascad**

# Execution report
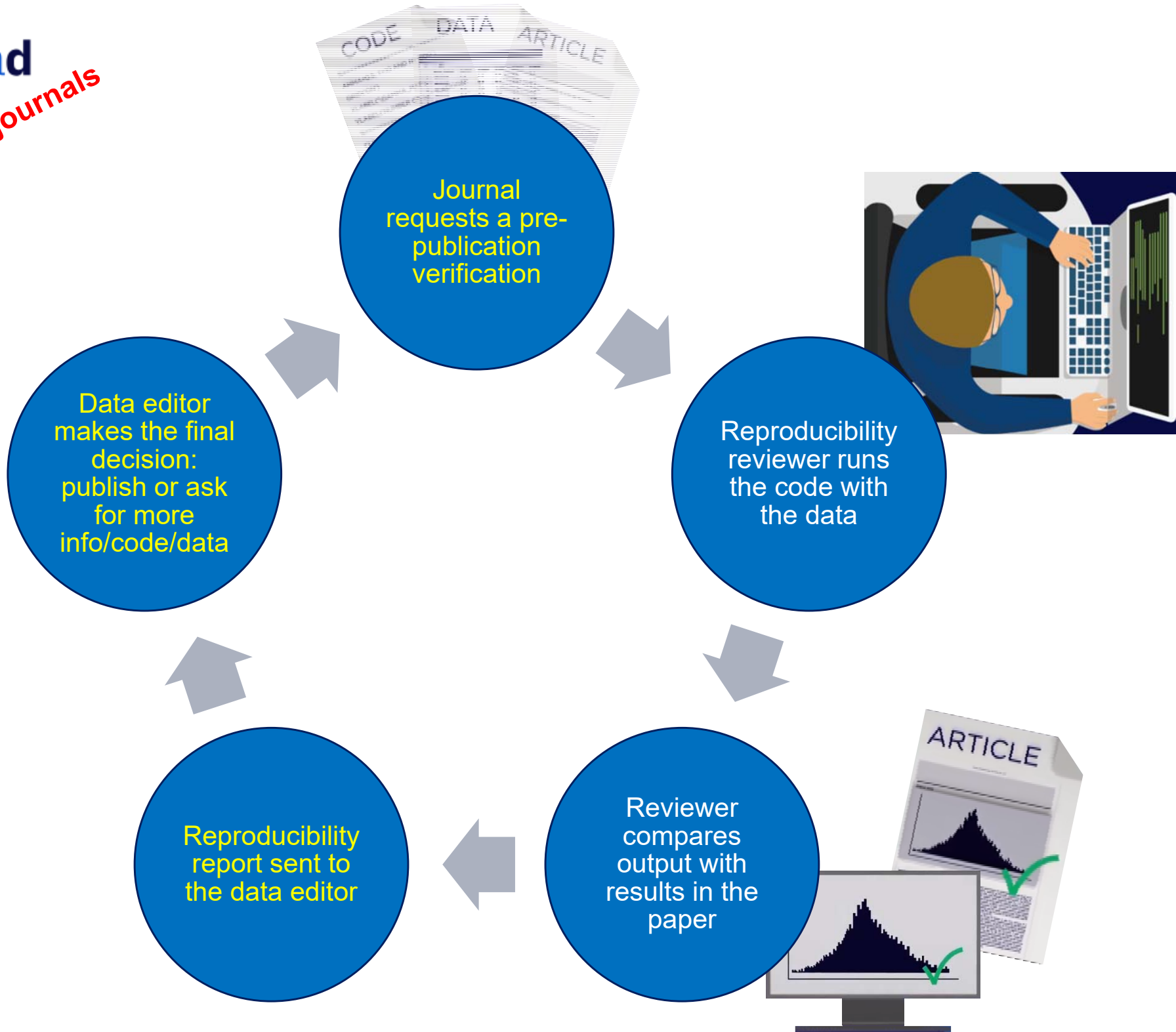
## 1. DATA DESCRIPTION

The results covered by the scope of this certification are generated using the French DADS (*Déclarations Annuelles des Données Sociales*, Annual Declaration of Social Data). Those datasets are produced by INSEE (Institut de la statistique et des études économiques) and the French Ministry of Finance, and are confidential. Access is remote through a secure access computer box (SD-Box) and is administred by the Centre d'Accès Sécurisé aux Données (CASD). More information on the data can be found here: https://www.casd.eu/en/source/dads-panel-with-matched-data-from-edp/. Access authorisation needs to be obtained by the Comité du Secret Statistique (https://www.comite-du-secret.fr/). The CASD statistics team which provides all info on the accessing procedure can be reached at service@casd.eu. For a thorough description of all the datasets used in the study, please refer to section I of the paper.

**cascad**

# Execution report

## 2. CODE DESCRIPTION

For the purpose of the certification, we aimed to check all the tables and figure displayed in the paper. The replication materials have the following structure:

- analysis code
    - do
    - matlab
- cleaning and creation code
- data_clean
- data_raw
- figures
- logs
- tables
- temp
    - bootstrap
        * demand_est
        * derivative
        * moments
        * resample_data
        * s2
        * supply_est

"cleaning and creation code" contains do-files that use the raw data stored in "data_raw", and create cleaned datafiles that are stored in the "data_clean" folder (the researchers have already provided the clean data). The Stata and Matlab files that are stored in "analysis code" then use those datasets to generate the tables and figures in the folder of the same name. The replication materials are provided along with an Excel file called "dataset_list", that lists all the datasets, be them raw or clean.

# cascad

# Execution report

## 3. VERIFICATION STEPS

The resources were downloaded from the openICPSR repository and run as per readme, using Stata 16 and Matlab 2019a on a DELL Optiplex 7060, 16GB RAM, intel® Core™ i7-7500U CPU @2.70-2.90GHz, and Windows 10 OS. Since several Matlab files rely on bash scripting that can only work in a Linux environment, we used a virtual machine with a Debian distribution, on which we installed Matlab 2019a. We allocated 12 GB of RAM to the virtual machine.

We first recreated the clean datasets using the raw data and the code stored in "cleaning-and-creation-code" and we had no problem. We then ran the code stored in "analysis-code" but we encountered an issue when running "B2_general_demand.m", the first file that uses bootstrapping. We got the following error:

```
Index in position 1 exceeds array bounds.

Error in B2_general_demand (line 8)
temp = temp{1,1};
```

This issue occurs at the beginning of the file:
```
temp = extractBetween(getenv('PBS_JOBID'),'[',']');
temp = temp{1,1};
```

Indeed, temp is a 0 by 1 cell with no value inside. When opened, it only diplays "val = ".

We assumed this error occurred because this file relies on the shell scripts and we didn't activate them before. They must be run in a terminal in Linux, with the command "qsub" (this is only mentioned in the shell scripts, and not in the readme). We had to install this command by running "sudo apt install gridengine-client" (again, there is no mention of this in the readme). But even with this, we were unable to use those scripts, as we got the following error:

*Unable to run job: unable to send message to qmaster using port 6444 on host "cascad": can't resolve host name*
*Exiting.*

As a result, we were unable to run any program that relies on bootstrapping and we had to stop our reproducibility attempt.

# Execution report

## TABLE 4: QUANTILES OF WAGE GROWTH

**Original :**

|  | Wage growth quantiles | | | | |
|---|---|---|---|---|---|
|  | (1) | (2) | (3) | (4) | (5) |
| $ICT_0$ | -.105*** | -.105*** | -.107*** | -.121*** | -.110*** |
|  | (.027) | (.018) | (.015) | (.018) | (.028) |
| Observations | 4,972 | 4,972 | 4,972 | 4,972 | 4,972 |

**Reproduced :**

|  | Wage growth quantiles | | | | |
|---|---|---|---|---|---|
|  | est1 | est2 | est3 | est4 | est5 |
| $ICT_0$ | -0.105*** | -0.105*** | -0.107*** | -0.121*** | -0.110*** |
|  | (0.025) | (0.017) | (0.015) | (0.016) | (0.031) |
| Observations | 4,972 | 4,972 | 4,972 | 4,972 | 4,972 |

# Collaboration with the American Economic Association (AEA)

Link: [here](here)

## Report by the AEA Data Editor

omic Association (AEA)
s to "design and oversee
rategy for archiving and
and promoting reproduc-
and Hoynes 2018). The
ata Editor (Vilhuber 2019)
lement that mission. Since
conducted comprehensive
ducibility checks for all
s, developed guidance for
with peers at societies and
and elsewhere.
ience from the first full year
rification, we implemented
ts in mid-2020. We pro-
lance to authors depositing
he data and code availabil-
and expanded and clarified
n third-party reproducibility
tion updates to replication
uired replication materials
experiments (Section I). We
cation reproducibility checks

replication package submissions we received. It
quickly became clear that stronger guidance and
greater clarity were needed to assist authors in
complying with the DCAP. Authors struggled
with how best to document their code and data;
the process of depositing data and code; and the
ability to provide clear data provenance, includ-
ing data citations.

To address these issues and improve compli-
ance upon submission, we took a two-pronged
approach: we clarified the policy and provided
improved instructions and guidance on how to
comply with the policy. We released the revised
version in September 2020. The main content
remains unchanged, but we simplified the main
policy, separated out the policy as applied to
papers conducting (field and lab) experiments,
and expanded the policy to encompass more
clearly any primary data collection. We also
introduced supplementary policies that lay out
how and when we expect reproducibility checks
by third parties to be conducted (also see our
interactions with third-party verifiers) and, as
a logical consequence of more transparent data

T BY THE AEA DATA EDITOR

*815*

earlier, 36 reports were provided by external
replicators or replication services for 27 man-
uscripts (see Table 2 for statistics by journal,
Appendix B for a list of third-party replicators).
Of those, several were provided by institutions
that already are organized as a reproducibility
service: Results Reproduction (R-Squared) at
Cornell University and cascad in France. We
appreciate the willingness of all third-party rep-
licators to provide us with their time and effort
in reproducing papers. In particular, cascad gen-
erously provided us with 21 reports. The AEA
data editor has had preliminary discussions with
several institutions about the interest and possi-
bilities of formalizing such services. Issues of
cost, frequency, speed, and at what point such
services would be involved in the research life
cycle remain unresolved.

### IV. Working with the Economics Community to Enhance and Broaden Education on Replicable Science

We have already noted the outreach to other
urnals and repositories above. Educating th
rent and f

**Other notable partnerships**

- **Econometrics Society:** **Econometrica**



- **Elsevier:** **International Journal of Forecasting**

- **Informs:** **Management Science**

Pérignon, Gadouche, Hurlin, Silberman and Debonnel (2019) Certify reproducibility with confidential data, *Science*
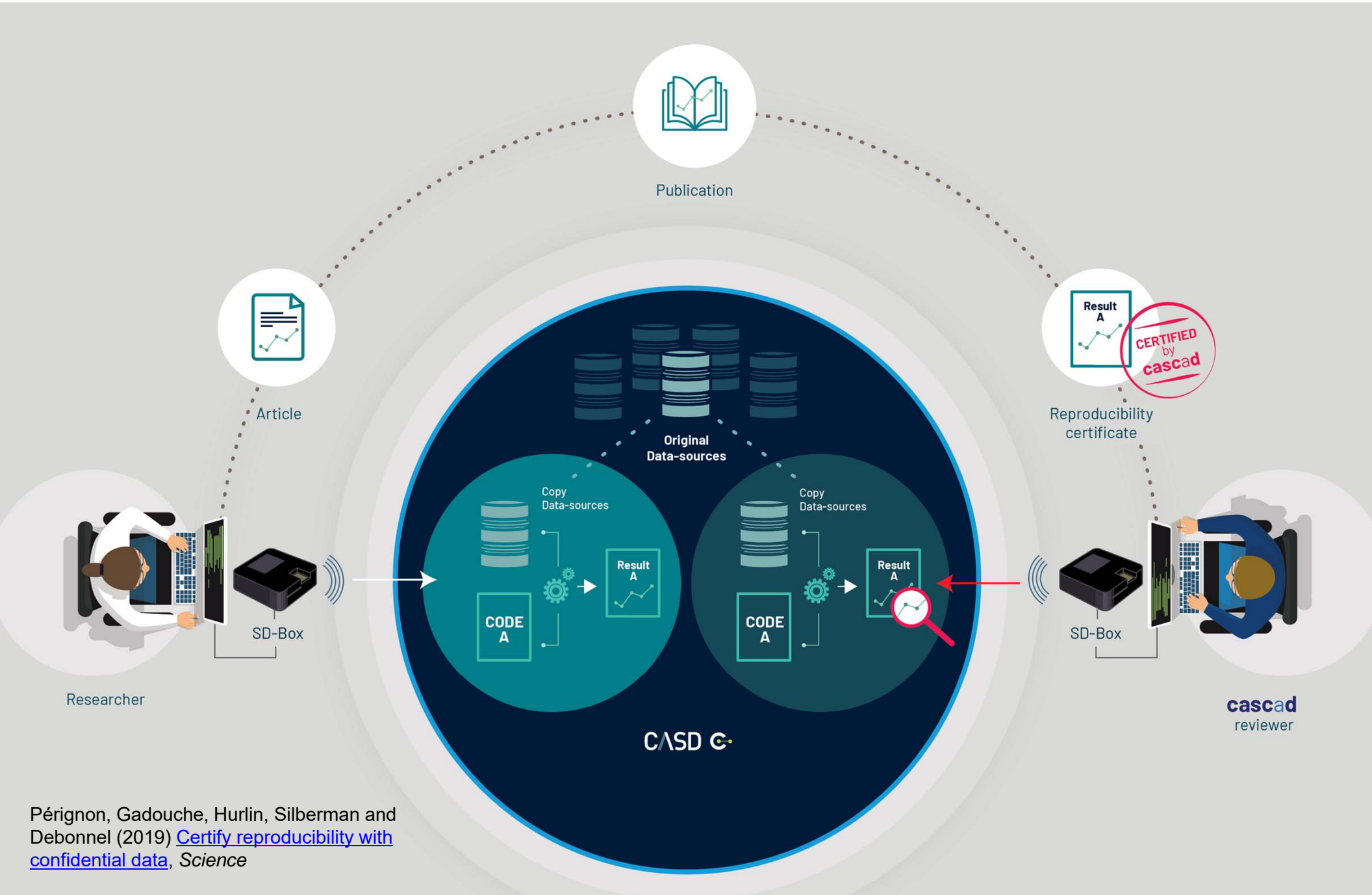
- Economies of scale (CPU, software, skills, access to data)

- Reduce conflicts of interest (journals, organisations)

- Can work in (virtually) any scientific area

- Detect problems early in case of pre-submission checks

- Act as an insurance policy for some authors

cascad

HDSR

📖 Issue 6.2, Spring 2024 ▾

# The Role of Third-Party Verification in Research Reproducibility

*by Christophe Pérignon*

**Column Editor's Note:** *Christophe Pérignon, in his contribution to this "Reinforcing Reproducibility and Replicability" column, describes how an innovative institution, called cascad, works. Pérignon, together with collaborator Christophe Hurlin, founded cascad to support researchers in verifying computational reproducibility before they would submit to journals. In contrast to Limor Peer, who described in a previous column how her institution provides such services for researchers affiliated with that institution, cascad is offering such reproducibility services to a broad audience in economics and management.*