

How to download datasets for total articles and researcher growth related to “The Strain on Scientific Publishing” by Hanson et al. (2023; arXiv).

## Contents

Scimago data download and assembly instructions .....	1
OECD, NSF (2022), Zwetsloot et al. (2021) PhD Data .....	2
UNESCO Researchers-per-million data .....	3

## Scimago data download and assembly instructions

1. Go to: <https://www.scimagojr.com/journalrank.php>
2. Download the data as per instructions in screenshot below. Note that the full dataset must be downloaded one-by-one as .csv files for each year. Place these in the folder **Data/raw\_data\_Scimago**

3. The script **0\_Merge\_raw\_Scimago\_data.R** will assemble all .csv files in **Data/raw\_data\_Scimago** into the final dataframe used in the analysis, saved as **Scimago\_data\_filtered.csv**.
4. As part of **Analysis.R**, there is a line in the script to run this as part of the overall script. Because this is a time-consuming data assembly. Check that a # is not hiding this script from being run. The relevant part of **Analysis.R** is:

```
##### 0. assemble Scimago annual .csv files into single dataframe #####
```

```
...
```

```
# source("Scripts/0_Merge_raw_Scimago_data.R")
```

Change the source(...) part to remove the hashtag:

```
source("Scripts/0_Merge_raw_Scimago_data.R")
```

5. Ready for Analysis.R! These data are used primarily for Figure 1, Figure 5, and related supplementary figures.

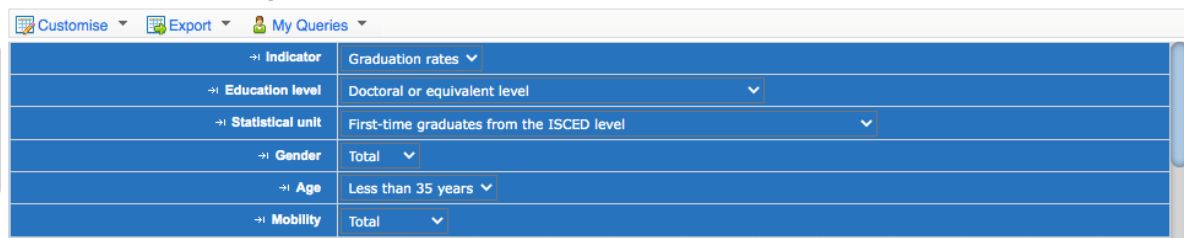
## OECD, NSF (2022), Zwetsloot et al. (2021) PhD Data

OECD data were downloaded per the following parameters in April 2023 from the link below:

[https://stats.oecd.org/Index.aspx?DataSetCode=EAG\\_GRAD\\_ENTR\\_RATES](https://stats.oecd.org/Index.aspx?DataSetCode=EAG_GRAD_ENTR_RATES)

This is a dataframe of graduation rates at the doctoral or equivalent level, including all first-time graduates from the ISCED level less than 35 years old. An analysis of these data has already been done by others, reported here: <https://master-academia.com/number-of-phds/>

### Graduation rates and entry rates <sup>+</sup>



Filter	Value
Indicator	Graduation rates
Education level	Doctoral or equivalent level
Statistical unit	First-time graduates from the ISCED level
Gender	Total
Age	Less than 35 years
Mobility	Total

Other datasets were included to supplement numbers for India and China for consideration in the supplemental figures. Because these are amalgamations across datasets, we preferred not to present the India- and China-supplemented data in the main text, since the underlying numbers across countries were generated through different methodologies. But like this, we intended to consider whether inclusion of India and/or China might change the overall picture of global PhD growth in a meaningful way, which it did not.

NSF (2022), used for data on India. Note, China used Zwetsloot et al. (2021) because Zwetsloot et al. (2021) provided a more up-to-date dataset and included projections into future years, which meant we did not have to model this independently: <https://nces.nsf.gov/pubs/nsb20223/figure/HED-29>

Zwetsloot et al. (2021), used for data on China: <https://cset.georgetown.edu/wp-content/uploads/China-is-Fast-Outpacing-U.S.-STEM-PhD-Growth.pdf>

These NSF and Zwetsloot et al. data were input directly into the R script **Fig1\_supp\_OECD.R** as a dataframe compiled from independent objects to make data sharing and assembly a tad easier.

# UNESCO Researchers-per-million data

Data downloaded from the UNESCO data portal:

[http://data.uis.unesco.org/Index.aspx?DataSetCode=DEMO\\_DS](http://data.uis.unesco.org/Index.aspx?DataSetCode=DEMO_DS)

1. UNESCO researchers-per-million data were downloaded July 2023 per the settings in the screenshot below. The data were saved as *UNESCO\_researchers\_per\_mil.csv* for easier processing:

The screenshot shows the UNESCO data portal interface. The main content area displays the 'Science, technology and innovation' dataset. The 'Indicator' dropdown is set to 'Researchers per million inhabitants (FTE)'. The 'Export' button is circled in red with the text 'Download here'. The table shows data for various countries from 2017 to 2022. Red annotations include 'Should say' pointing to the indicator name and 'Tick this box' pointing to a checkbox in the left sidebar.

Country	2017	2018	2019	2020	2021	2022
Albania	...	832.42282	...	...	...	...
Angola	18.9549	...	...	...	...	...
Argentina	1 280.70106	1 212.45726	1 227.40422	1 231.51703	1 256.26717	1 283.79869
Armenia	...	...	...	...	...	1 219.93327
Australia	...	...	...	...	...	...
Austria	5 401.30362	5 416.8937	5 683.24278	5 959.54993	5 829.83733	6 341.73871
Azerbaijan	...	...	1 694.46991	1 712.16743	1 728.76393	1 690.6737

2. **Some caveats to this dataset and its use:** researchers-per-million reporting is spotty or absent from year to year for some important countries. For instance, the following countries provide no data, or have absent data from 2019-on: Canada, New Zealand, Switzerland, India. From 2020-on, additionally the United States of America and China were lacking data at the time of download, and the United Kingdom lacked data but provided an estimate.
3. **Because of caveats listed in #2:** collecting data on researchers-per-million at the world level may give a skewed representation of the underlying data as this number is not informed in 2019/2020-on by countries where data are absent.