

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Survey was run and its data was collected on Qualtrics. The gender of the authors in the dataset, was collected from the API of Genderize.io.

Data analysis No software was used to analyze the data, we just used python programming language.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The details of all the data and methods used to generate the data are provided in the Methods section and Supplementary Information. The source data for all figures is also provided. The following data availability statement is included: "Source data for all figures in the main manuscript and the supplementary are provided with this paper. Microsoft Academic Graph (MAG) can be downloaded from <https://bit.ly/3kPaUqe>. All data generated from MAG for the purpose of this study is made available at <https://bit.ly/3cHJJuC>. A reporting summary for this Article is available as a Supplementary Information file."

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We use Coarsened Exact Matching to see if mentorship in academia has a causal effect on a scientist's publication impact. The data is quantitative.
Research sample	The dataset was downloaded from Microsoft Academic Graph's website. It includes information on 215 million scientists and 222 million publications. The dataset includes detailed records of scientific publications and their citation network. The whole dataset was used during the analysis to identify 3 million mentor-protege pairs for the study.
Sampling strategy	This is not applicable for the main study. For the survey, a random sample of 2000 scientists was selected and requested to fill out our survey. We focused on scientists who completed their mentorship period in the recent years to minimize recall bias in their responses, and maximize out chances of finding an active affiliation for them engaged as researchers. The survey aimed to explore the lived experiences of scientists in collaborations and mentorship relationships.
Data collection	The data for the main study was downloaded from Microsoft Academic Graph's website on Jan 2020. The data collected for the survey was collected over a period of 4 weeks from February 19th to April 28th 2020. The survey did not use randomization, it is not an experiment. The data were collected using Qualtrics, and email addresses of scientists were retrieved from publicly available sources, such as university websites. The gender of the scientist was generated from a gender classification tool: genderize.io.
Timing	The data for the main study was downloaded from Microsoft Academic Graph's website on Jan 2020. The data collected for the survey was collected from February 19th to April 28th 2020. The survey was created at this time in response to reviewers comments and suggestions.
Data exclusions	As we are studying the effect of mentorship in this study, any scientist that was not identified as a protege or a mentor (as per our pre-defined criteria) was excluded from the analysis. The scientists contacted for the survey were not included in the analyses if they have not recognized the collaboration or the collaborator we asked them about.
Non-participation	Participants were only used when conducting the survey. Out of the 2000 scientists we emailed, 167 participated in our survey by filling it out completely, and having recognized the collaboration and the collaborator they were asked about.
Randomization	For the survey, a random sample of 2000 scientists were selected from our dataset of 3 million. We focused on scientists who completed their mentorship period in the recent years to minimize recall bias in there responses, and maximize out chances of finding an active affiliation for them engaged as researchers.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

Recruitment

Study participants in the survey were invited to participate via email. The data were collected using Qualtrics, and email addresses of scientists were retrieved from publicly available sources, such as university websites. The survey used incentives, and extended a \$10 Amazon gift cards available on 5 different Amazon sites of the respondent's choosing. Respondents were allowed to forego their fixed reward and enter in a raffle for 10 \$100 Amazon gift cards, which we distributed to randomly selected participants who chose this option. Participants were not informed of the nature of the research question, namely, that we were studying the relationship between mentorship quality and scientific output post mentorship to make sure that participants do not select into the survey on the basis of the quality of mentorship they perceive to have received. We analyzed survey drop-out and have not identified any given point where the few respondents who left the survey after clicking on the survey link would leave. In sum we do not believe that the sample is biased when it comes to the relationship we set out to test.

Ethics oversight

The survey portion of the study was approved by the NYUAD Institutional Review Board, #HRPP-2020-8.

Note that full information on the approval of the study protocol must also be provided in the manuscript.