Who does big team science?

Erin M. Buchanan[1] & Savannah C. Lewis[2]

[1] Harrisburg University of Science and Technology

[2] University of Alabama

Author Note

Erin M. Buchanan is a Professor of Cognitive Analytics at Harrisburg University of Science and Technology. Savannah C. Lewis is a graduate student at the University of Alabama.

Thank you to Dwayne Lieck for providing an extensive list of large scale projects for this manuscript.

The authors made the following contributions. Erin M. Buchanan: Conceptualization, Data curation, Formal Analysis, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing; Savannah C. Lewis: Conceptualization, Data curation, Methodology, Project administration, Writing – original draft, Writing – review & editing.

Correspondence concerning this article should be addressed to Erin M. Buchanan, 326 Market St., Harrisburg, PA 17101. E-mail: ebuchanan@harrisburgu.edu

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: big team, science, authorship, credit

Word count: X

Who does big team science?

The introduction will go here. Here's an outline:

- Big Team Science
  - one off papers
  - collaborative teams
- Credibility revolution
- WEIRD
- ... more tbd, brain isn't braining
- Research Question 1: What journals publish big team science papers?
- Research Question 2: What are the types of articles that are being published in big team science?
- Research Question 3: Who is involved in big team science?

For each of these research questions, we will examine the overall results of all big team research projects, and examine for change in result trends across years of publication.

Method

Studies

We defined **big team science publications** as publications with at least 10 authors that were published in peer-reviewed journals or had posted a full paper pre-print for publication review. We specifically focused on social science research, primarily *psychology* for this manuscript. First, we added all known publications from collaborative teams, such as the PSA, Many Labs, and Many Babies. We examined journals that frequently publish

registered replication reports (i.e., *Advances in Methods and Practices in Psychological Science*) for additional publications with at least 10 authors. From these manuscripts, we identified common authors who frequently participate in these studies, and examined their Google Scholar or Open Researcher and Contributor IDentifier (ORCID) page for other publications. We reached out to social networks on Twitter to identify other publications. Last, we used Google Scholar and EBSCO to search for large projects using the following search terms: collaboration, multicultural, large scale, and big team science.

Data Curation

RQ1: Journal Information. Using these criteria, we identified 70 articles for inclusion on this manuscript. The publication dates on these articles ranged from 2013 to 2022, and we used the pre-print last updated date as the publication date for those articles. The current impact factor (i.e., 2022) for each journal was found on the journal page and included for journal statistics.

RQ2: Article Information. For each publication, we coded the list of keywords into broad labels for areas of social science (i.e., Social Psychology, Cognitive Psychology). In this section, we elected to code each article with only one main research area. Research is often cross-disciplinary, however, for simplicity, we applied one global label to each article that represented the perceived main area of study (based on what course this material might be taught in, the publication journal focus). The list of broad areas includes clinical, cognitive, developmental, educational, and social psychology. Last, we included a metascience category that covered articles that were detailing the science of science. The Open Science Collaboration [CITE] was included in this category because the aim of the paper was a focus on replication across multiple field types. Other metascience categories included papers that focused on research degrees of freedom, analysis choices, and types of samples.

RQ3: Author Information. The author list was then extracted from each publication. In the case of consortium authorship, we extracted the complete authorship from the meta-data or pre-print publication. The total number of unique authors was 3336. The number of authors on each publication ranged from 1 to 482 with an average of 68.63 authors ($SD = 89.89$).

Next, we matched each author to their Google Scholar and ORCID profile pages, if available. We originally used the *R* packages, *rorcid* [CITE] and *scholar* [CITE] to try to match published author names to profile pages. This process did not result in a large number of matches, and we therefore curated the list of profile pages manually, checking each author against the publication list. We used these two packages and profile pages to collect authorship statistics described below.

Career Length. Career length for each author was defined using multiple variables to see if results from the two data sources would converge on similar answers. Both ORCID and Google Scholar provide a list of publications for authors, and we first calculated career length as the year of first publication listed for each author. In ORCID, a researcher can enter their educational background with completion years for each degree. We defined career length for this variable as years since first degree listed. Publication years are often curated directly from meta-data provided by Crossref (ORCID) or online sources used by Google Scholar. Authors may also directly add publications and their information into both systems. The limitation to using education as a metric for career length is that the researcher must directly enter this information into ORCID.

Employment. Employment information was collected from self-entered ORCID data. These values are open text, and therefore, we coded them into coherent categories for traditional educational (graduate student, post doctoral, lecturer), tenure track (assistant, associate, full professor), and other roles (fellow, research assistant, researcher, head). Employment geopolitical region was also selected when available.

Education. As with employment information, we also collected education information from self-entered ORCID data. These values were coded into general categories of bachelor, master, and doctoral degrees. The geopolitical region of the listed education was included when available. For analyses, both employment and education levels were grouped into United Nation regions.

Types of Publications. ORCID includes information about the type of publication pulled from either researcher entered data or Crossref. We coded these publications into general categories including book, conference presentations, data-sets, journal articles, preprints, software, thesis, and other publications.

Publication Metrics. We calculated total number of publications of any type from both Google Scholar and ORCID. We additionally pulled both the h-index and i-10 index from Google Scholar. The h-index represents the highest h number of publications that have at least h citations, while the i-10 index represents the number of publications with at least 10 citations.

Data analysis

RQ1: Journals. Results for types of journals will include a summary of the journals that publish big-team science papers and an average of the 2-year and 5-year impact factors for the journals.

RQ2: Articles. Results for types of articles will include a summary of the coded areas for each article, presented overall and across the years of publication.

RQ3: Authors.

We will use $\alpha < .05$ for all analyses that involve hypothesis testing. We make no directional predictions.

Career Length. For each of the three variables in career length (ORCID year of first pub, Scholar year of first pub, ORCID year of first degree), we will create a

visualization of the trend and variance of researcher career length across publication years. To analyze trends over time, we will use the slope from a multilevel model (MLM) using the individual as a random intercept, career length as the dependent variable, and year of publication as the predictor variable. MLMs are regression models that control for the correlated error due to the repeated and nested nature of the data [CITE]. In this model, each individual's starting career length is arbitrary, which we allow to vary with the random intercept by participant. A positive slope for year of publication would indicate increasing years of first publication (i.e., more younger scholars over time), while a negative slope would indicate older years of first publication (i.e., more older scholars over time). In order to show variance between individuals, we will report the variance component of the random intercept for individuals. Finally, to examine variance over time, we will calculate the standard deviation of the individual within each publication (i.e., one standard deviation for each article), and use a linear model with year predicting these standard deviations. This model is a traditional simultaneous regression, as averaging the variance by title eliminates the repeated measures variable (individuals). A positive slope would indicate increasing variance over time (i.e., more diversity in the career lengths of scholars), while a negative slope would indicate less variance and diversity in scholars over time.

Employment. We will summarize the general employment categories of individuals at the time of publication, along with a summary for change over time.

Education. We will summarize the general education categories of individuals at the time of publication, along with a summary for change over time.

Types of Publications. We will summarize the coded types of publications for individuals.

Publication Metrics. We will report descriptive statistics on the total number of publications, i10 index, and h-index for individuals overall. Next, we will use the same

analyses described in the career length section to analyze trends over time. An increasing slope over time indicates that individuals who are publishing more are more represented in big-team science over time (i.e., increasing numbers of scholars with higher publication rates), while a negative slope indicates more researchers with less publications. A positive slope for variance indicates increasing variance over time (i.e., more diversity in the individual publication rates), while a negative slope would indicate less diversity in researchers over time. While publication rates do not represent value as a researcher, they are often used in hiring and promotion decisions, and we will use this variable as a proxy to gauge the diversity in scholars represented in big teams.

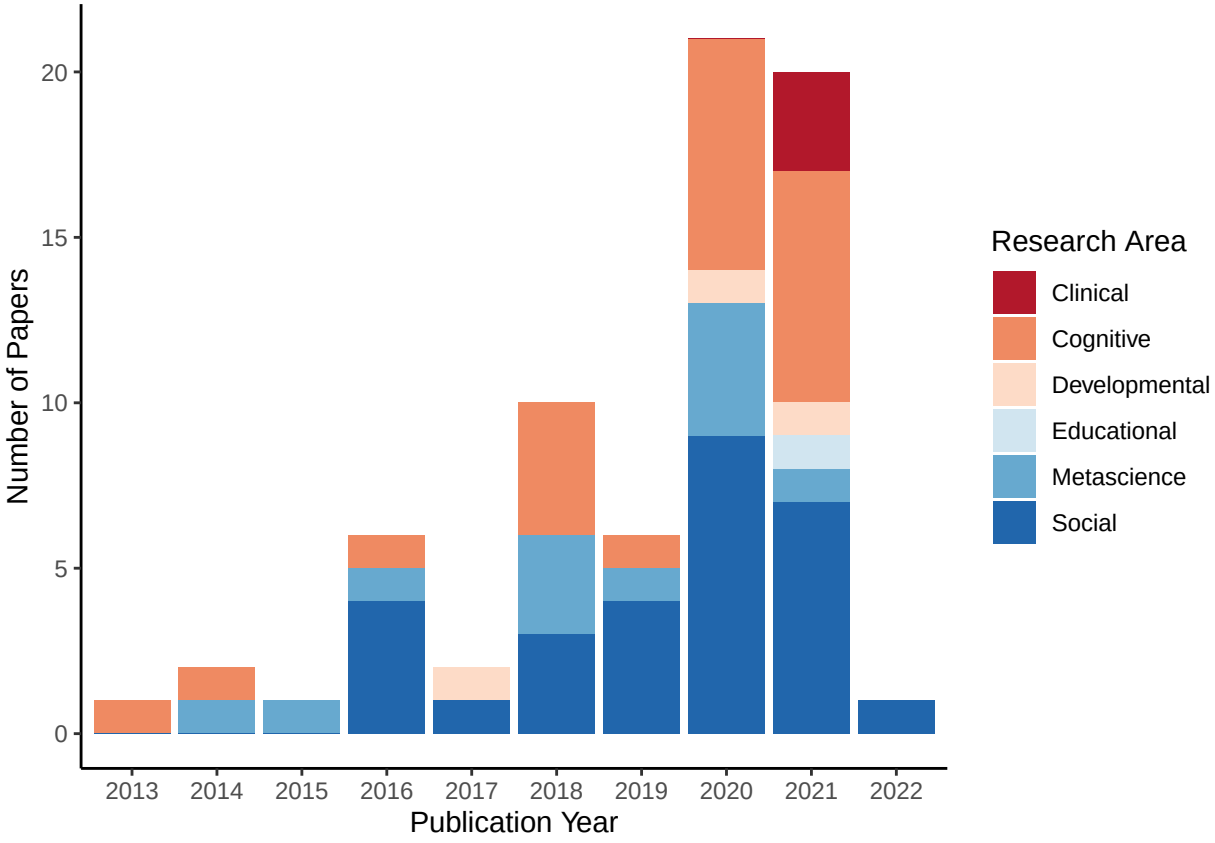
Geopolitical Regions. We will present visualizations of the last listed education and job location, and we will discuss the areas of world in which authors generally come from, as well as the lowest representation of authors. To understand the change in representation diversity, we will summarize the total number of geopolitical regions for each paper. Using a linear model, we will examine if the number of regions present is predicted by the year of publication. Increasing diversity would be represented by a positive slope, while decreasing diversity would be represented by a negative slope. Last, we will examine the differences in representation for corresponding author sets versus all other authors. For papers with 10 to 49 authors, we will use the three first authors and the last author to compare against random samples of other authors. For 50 to 99 authors, five first authors plus last will be used, and for all papers with more than 100 authors, we will use ten first authors and the last author. We will calculate the frequencies of each of the UN Sub-Regions for first authors. Then we will randomly sample all authors in subsets of the same size (i.e., 3+1, 5+1, 10+1), and average the frequencies of UN Sub-Regions across these random samples. We will randomly sample the same number of times as the size of the author list (i.e., a paper that has 54 authors will be randomly sampled 54 times with a sample size of six).

Given the expected small sample sizes of these contingency tables (i.e., the largest could be $n = 22$ if all author information is available), we will group together titles based on time period using pre-2020, 2020, and post-2020 to show change over time. The choice of these groupings is based on the number of articles avaliable within these time brackets. For each grouping, we will calculate the effect size of the differences in frequencies comparing first authors to all other authors. Since this data is categorical, we will use Cramer's V to represent the effect size. If the effect size includes zero in it's confidence interval, this result will imply that first and all other authors represent the same pattern of UN Sub-Region diversity. Any confidence interval that does include zero represents a difference in diversity. We will report these values and discuss what regions of the world are represented when effect sizes indicate a different from zero using standardized residuals.

Results

RQ1: Journals. Articles were most commonly published in Advances in Methods and Practices in Psychological Science ($n = 20$), Pre-Print ($n = 17$), Perspectives on Psychological Science ($n = 7$), PLOS ONE ($n = 4$), and Nature Human Behaviour ($n = 3$). A complete list of journals can be found on our Open Science Framework page XXX. The average 2-year impact factor for official journal publications was 8.22 ($SD = 10.67$) and the average 5-year impact factor was 6.41 ($SD = 4.42$).

RQ2: Articles. Articles were primarily Social (41.4%) and Cognitive (31.4%), followed by smaller categories for Metascience (17.1%), Clinical (4.3%), Developmental (4.3%), and Educational (1.4%). Here, we will talk about the figure below and how the trends occur across time.



RQ3: Authors.

Career Length. The average career length was 11.97 ($SD = 7.95$) for ORCID first publication year (i.e., current year minus year of first publication), $M = 13.12$ ($SD = 8.74$) for Scholar year of first publication, and $M = 13.54$ ($SD = 7.77$) for ORCID year of first degree. In each of the analyses, we use the actual year of first publication/education, as subtracting publication and using it as a predictor would create a perfect solution for each model.

Next, we will talk about the results from three variables:

1) ORCID year first publication:

- The slope for year of first publication was $b = 0.00$, $SE = 0.00$, $t(53) = 1.71$, $p = .092$.
- The variance parameter was $SD = 7.99$.
- The slope for variance across years was $b = 0.01$, $SE = 0.48$, $t(23) = 0.02$, $p = .983$.

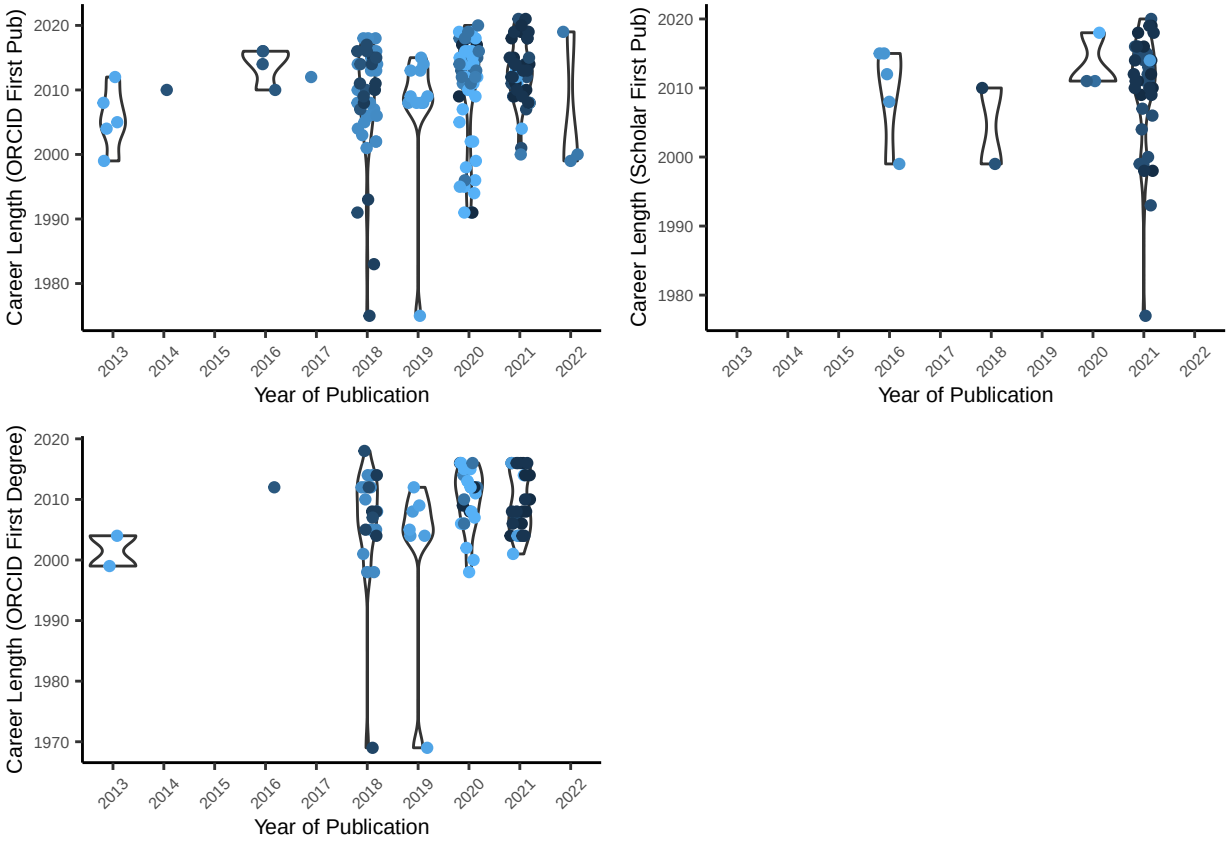
2) Scholar year first publication:

- The slope for year of first publication was $b = 0.00$, $SE = 0.00$, $t(15) = 1.59$, $p = .132$.
- The variance parameter was $SD = 8.88$.
- The slope for variance across years was $b = 0.16$, $SE = 0.61$, $t(2) = 0.26$, $p = .821$.

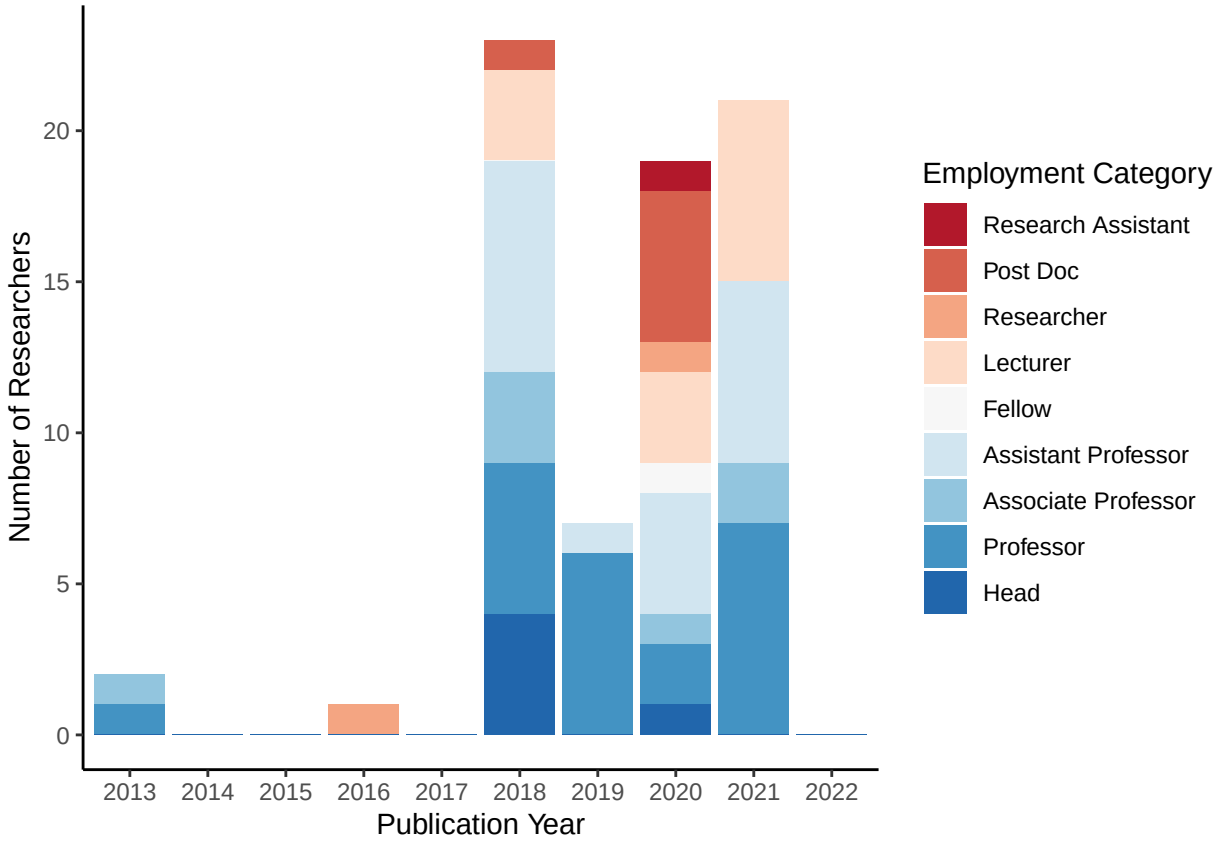
3) ORCID year first degree:

- The slope for year of first degree was $b = 0.00$, $SE = 0.00$, $t(35) = 5.14$, $p = < .001$.
- The variance parameter was $SD = 7.77$.
- The slope for variance across years was $b = 0.03$, $SE = 0.82$, $t(14) = 0.04$, $p = .971$.

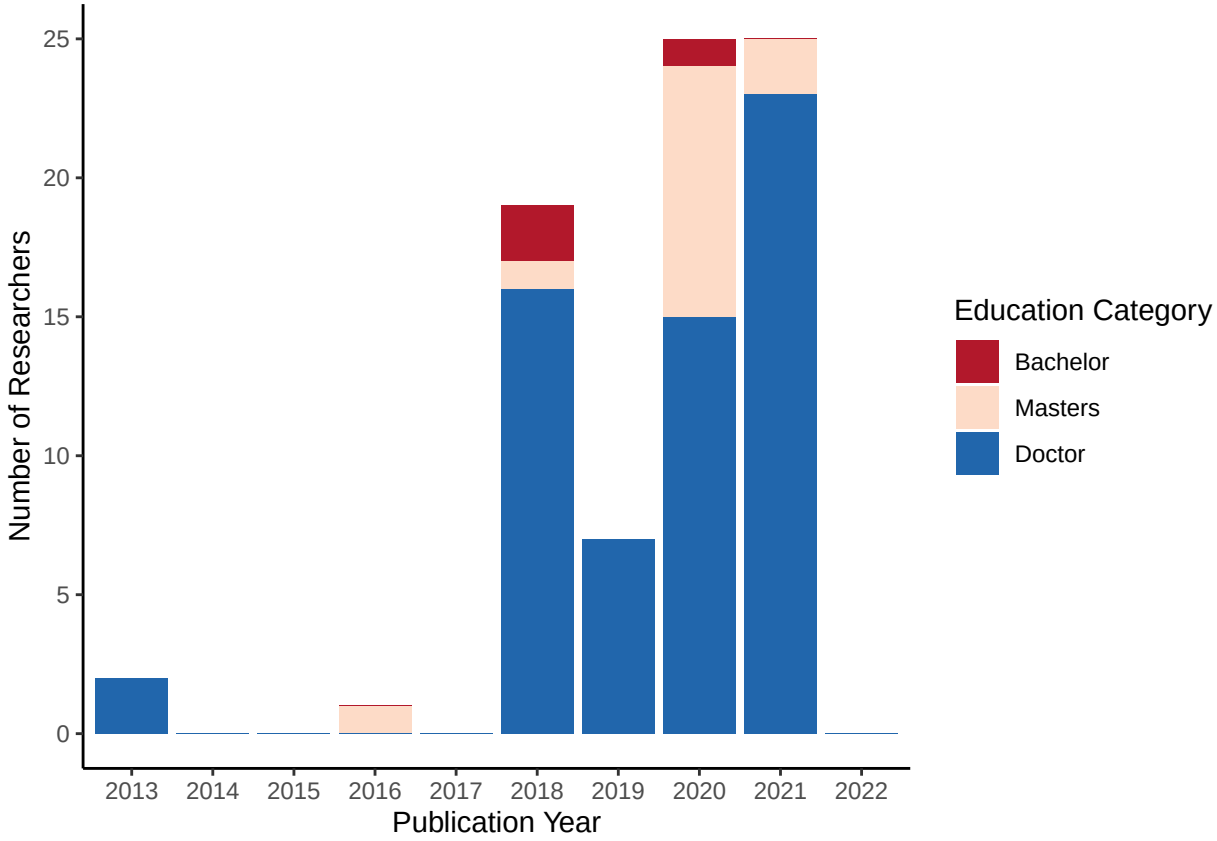
Then we will discuss how these results are represented in our graph (below). We will discuss any discrepancies in the results across the three different statistics.



Employment. Overall, individuals **at the time of publication** in big team science publications are Professor (28.8%) and Assistant Professor (24.7%), followed by smaller categories for Lecturer (16.4%), Associate Professor (9.6%), Post Doc (8.2%), and Head (6.8%). Here, we will talk about the figure below and how the trends occur across time. We will also comment on the number of jobs that would be considered "non-academic".



Education. Overall, individuals **at the time of publication** in big team science publications are Doctor (79.7%), Masters (16.5%), and Bachelor (3.8%). Here, we will talk about the figure below and how the trends occur across time.



Types of Publications. In general, the most common type of publication was journal-article (78.7%), followed by other (7.7%), and smaller categories for conference (5.4%), book (5.1%), preprint (1.3%), and software (0.7%). We will include all information over 1% of the publication totals.

Publication Metrics. The average number of all publications per individual was M = 52.69 (SD = 70.37) using ORCID and M = 64.20 (SD = 110.06) using Google Scholar. The average i10 index was M = 30.96 (SD = 56.44) and the average h-index was M = 18.55 (SD = 21.89).

Using the same analyses as career length, we will report:

1) Number of publications ORCID:

- The slope for number of publications was b = -0.02, SE = 2.02, $t(22)$ = -0.01, p =

.991.

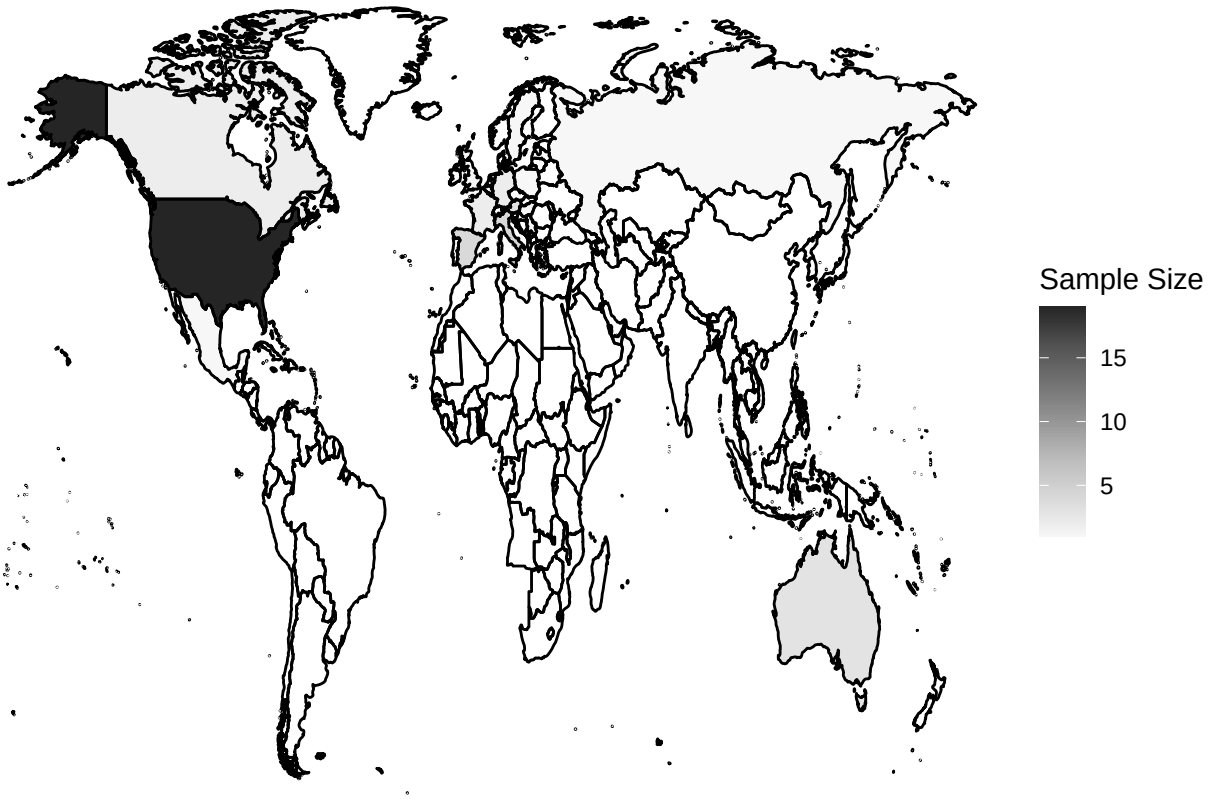
- The variance parameter was $SD = 29.13$.
- The slope for variance across years was $b = 1.94$, $SE = 5.44$, $t(23) = 0.36$, $p = .725$.

2) The number of publications for Scholar:

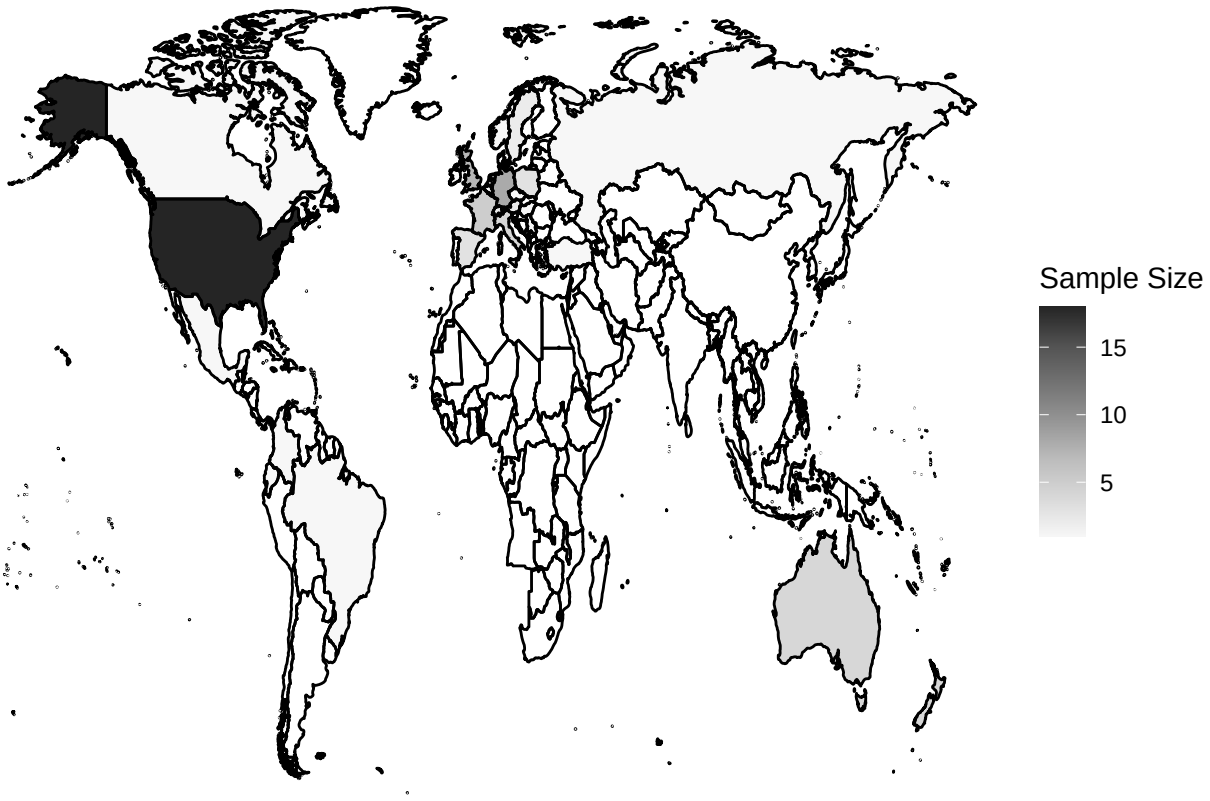
- The slope for year of first publication was $b = 0.00$, $SE = 0.00$, $t(15) = 1.68$, $p = .114$.
- The variance parameter was $SD = 100.62$.
- The slope for variance across years was $b = -38.52$, $SE = 6.27$, $t(2) = -6.14$, $p = .026$.

Geopolitical Regions. We will first present two figures for education and employment geopolitical regions representation in scholars.

Education Graph



Employment Graph



Next, we will report the results of two linear regression models. The independent variable is year of publication, and the dependent variables include the summary of the total number of different geopolitical regions represented in each publication. - The slope for the education model was $b = 0.06$, $SE = 0.09$, $t(68) = 0.63$, $p = .531$.

- The slope for the employment model was $b = 0.01$, $SE = 0.09$, $t(68) = 0.12$, $p = .903$.

Last, we will examine the effect sizes of the UN Sub-Regions represented for each grouped year of analysis. Tables of the UN Sub-Regions will be presented in our supplemental documents.

- Pre-2020: $V = .36$, 95% CI [.54, .78] and 0.13, 0.44, 0.94, -0.57, -0.84, -0.57, -0.13, -0.44, -0.94, 0.57, 0.84, and 0.57 will be used to discuss pre-2020 manuscripts.

- 2020: $V = .64$, 95% CI [.48, 1.19] and -0.66, 2.31, -1.02, -0.84, 0.66, -2.31, 1.02, and 0.84 will be used to discuss 2020 manuscripts.

- Post-2020: $V = .55$, 95% CI [.65, 1.28] and 1.45, -0.68, -0.68, -0.68, -1.45, 0.68, 0.68, and 0.68 will be used to discuss post-2020 manuscripts.

Discussion

To be included after we have completed analyses.

Limitations

- Discuss the limitation of "who is using ORCID and Scholar" which can bias our results.
- Other limitations will be included.

283 ## References