Supplement to Bruna: Fundamental errors of data collection & validation undermine claims of 'Ideological Intensification' made by the National Association of Scholars

Emilio M. Bruna[1,2]

[1] Department of Wildlife Ecology and Conservation, University of Florida, PO Box 110430, Gainesville, FL 32611-0430, USA

[2] Center for Latin American Studies, University of Florida, PO Box 115530, Gainesville, FL 32611-5530, USA

Author Note

All code and data used in this analysis are available at https://github.com/embruna/quantdei_nas.

Data Review and Validation

The search for duplicated records and other validation procedures were carried out using code written in the R statistical programming language (*1*) with functions from the `tidyverse` (*2*) and `janitor` (*3*) libraries. This code was then applied to three of Goad and Chartwell's `'clean'` data sets, all of which are located in subfolders of their Github repository's `'out'` folder (*4*):

1. University Twitter accounts: `tweets_clean.csv`
2. Research grants
    A. National Science Foundation (i.e., NSF): `nsf_all_grants_summary_data.csv`
    B. National Institutes of Health (i.e., NIH): `nih_parsed_all.fst`
3. Scientific publications in Google Scholar: `google_scholar.fst`

The code used to validate data and the resulting output are available at (*5*); below I provide summaries and representative examples of the errors revealed by the validation procedure. It is important to emphasize that any error estimates presented are conservative, as the validations carried out were merely a "first pass" using simple search strings. More robust validation efforts will almost certainly identify additional errors.

University Twitter accounts

Goad and Chartwell searched 895 university accounts for 21 terms they define as DEI-related (e.g., "advocacy", "ally", "diversity", "equity", "justice", "privilege", "race"). This resulted in 151284 tweets, which they then used to graph the use of the individual terms over time. Many of the terms for which they searched, however, also have uses and meanings beyond DEI. For instance, "race" could refer to competitions or athletic events, "ally" is a common nickname for "Allison", and introductions are often prefaced by the phrase "it is my privilege to...". Goad and Chartwell clearly failed to filter their dataset for tweets using these terms in non-DEI contexts; based on my preliminary review at least

6.78% of the tweets in their data set are not actually DEI-related, with the percentage of irrelevant tweets for a given term ranging from 0.56 - 39.51%.

NIH and NSF grants

Goad and Chartwell also failed to screen for alternative uses of their focal terms when reviewing the grants awarded by NSF and NIH (e.g., N = 2783 of the NSF grants they identify using the term "diversity" in a DEI-context are actually investigating genetic, phylogenetic, or species diversity. However, a more serious issue is that that they vastly inflated their sample sizes. When researchers at multiple institutions are involved in a project, they submit a single grant proposal. If selected for funding, the NSF and NIH will allocate each institution their portion of the grant funds directly. By failing to consolidate the awards for each of the co-PIs collaborating on the same grant proposal, Goad and Chartwell overstimated the number of NSF and NIH grants by at least 12.13% and 66.67%, respectively.

Scientific publications in Google Scholar

Finally, Goad and Chartwell sought to identify any DEI-related publications in the scientific literature. To do so they searched a number of repositories, including Google Scholar, for DEI-related articles in science, technology, engineering, and mathematics (STEM) journals by using search strings including a STEM-term and one of their DEI-related terms (e.g., "biology diversity"). Here again they failed to search their results for duplicate or irrelevant records prior to graphing their results - over 3550 of the records in their data set were duplicates (27.17%), and at least XXXX of the articles they considered DEI-related were incorrectly included. Moreover, their data set of 'DEI-related articles in STEM journals' included at least N = 473 articles in humanities, cultural studies, and law journals. A partial list of these journals can be found in Table SX.

References

1. R Core Team, "R: A language and environment for statistical computing" (manual, Vienna, Austria, 2020), (available at https://www.R-project.org/).

2. H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the tidyverse. *Journal of Open Source Software.* **4**, 1686 (2019).

3. S. Firke, "Janitor: Simple tools for examining and cleaning dirty data" (manual, 2021), (available at https://CRAN.R-project.org/package=janitor).

4. N. A. of Scholars, Quantitative Study of Diversity, Equity and Inclusion in STEM Subjects in US Universities (2022), doi:10.5281/zenodo.6360904.

5. BrunaLab/quantdei_nas, (available at https://github.com/BrunaLab/quantdei_nas).

Table 1

Irrelevant tweets attributed to seven different DEI terms and the total number of tweets for each term in the original dataset. Note that this percentage is a conservative estimate, as it is based on a preliminary review.

DEI Term	Irrelevant Tweets (N)	Total Tweets (N)	% Irrelevant
diversity	174	31268	0.56
equity	270	16374	1.65
ally	377	6953	5.42
advocacy	512	5128	9.98
justice	2491	22090	11.28
race	5051	25167	20.07
privilege	1382	3498	39.51

Table 2
Sample tweets incorrectly attributed to different DEI terms.

DEI Term	sample irrelevant tweet
advocacy	a passionate physician and educator commited ot medical education pati
advocacy	the basic trial advocacy class at the uarizonalaw school argued their
advocacy	students in the basic trial advocacy class at the uarizonalaw had thei
ally	allymahoney11 y grades can be given for students if the faculty member
ally	allyrae12508 congratulations
ally	allyrae12508 welcome to the sun devil family
diversity	a new university of arizonaled study uses big data to assess why the d
diversity	a new study coauthored by university of arizona researchers provides t
diversity	new uarizona research finds that sexual reproduction and multicellular
equity	judge rakoffs decision has the potential of really blowing up said bri
equity	highly speculative prof renee jones talks to businessinsider about pri
equity	rt hnbayld congrats alex mancebo jonesday boston office focuses his pr
justice	arizonapbs will honor the legacy of supreme court justice sandra day o
justice	asucrimjustice researchers have found that there is a higher likelihoo
justice	two weeks before her first year at asu carson swisher changed her majo
privilege	rt azathletics 90 years ago today beardown was born it is a privilege
privilege	rt uapolicechief thank u asuatoday amp uaaa for this very special hono
privilege	rt wendelldneal two of the greatest guys i have ever had the privilege
race	join the beantowncats for the jeff coombs memorial virtual road race a
race	ronald a wilson ua title ix director and a former presiding judge for
race	join the beantowncats in the jeff coombs memorial road race on sept 9

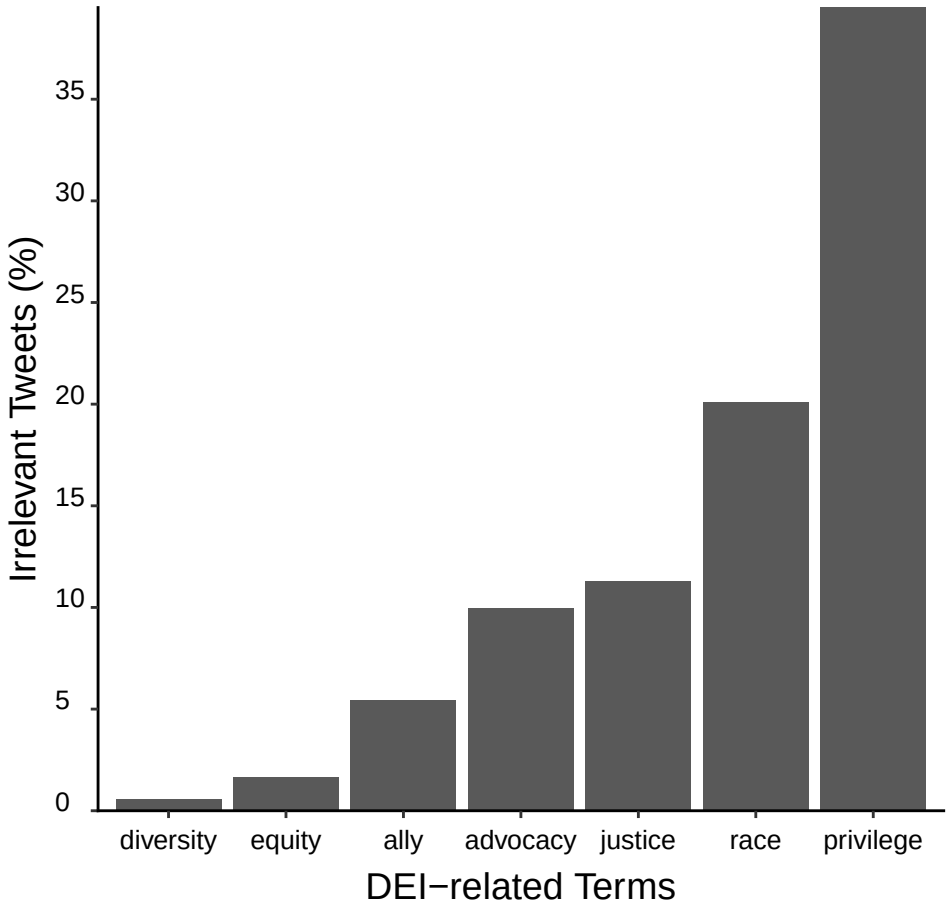

Figure 1. Percentage of irrelevant tweets atrtibuted to seven different DEI search terms. Note that this percentage is a conservative estimate, as it is based on a preliminary review.