Supplement to '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

Below I present a brief overview of the methods used to review the contents of 5 data sets used by Goad and Chartwell to visualize trends in DEI-language use. These data sets can be found in the `'out/twitter'`,`'out/grants'`, and `'out/scholarship'` folders of the NAS Report's Github repository (*1*).

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

Although many of these errors would be detected immediately by simply scanning the data sets, I wrote code in the R statistical programming language (*2*) to conduct some simple data validation tests. This code, which included functions from the `tidyverse` (*3*) and `janitor` (*4*) libraries for filtering, de-duplicating, and summarizing data frames, is available at (*5*), as are `.csv` files of the resulting output. Below I provide summaries and representative examples of the errors revealed by the validation procedures .

University Twitter accounts

Goad and Chartwell searched 895 university accounts for over 20 terms they define as DEI-related (*6*). They used the resulting dataset of N = 151284 tweets (`'tweets_clean.csv'`) to graph the use of the DEI-terms over time. Many of the terms for which they searched, however, have uses and meanings beyond DEI. For instance, "race" could refer to competitions or athletic events, "ally" is a common nickname for "Allison", "justice" is the title used by members of federal or state bench, and introductions are often prefaced by the phrase "it is my privilege to. . . ".

I reviewed Goad and Chartwell's twitter dataset for tweets that might be using seven of

their DEI-related search terms in a non-DEI context. These terms were: “advocacy”, “ally”, “diversity”, “equity”, “justice”, “privilege”, and “race”. I first filtered `'tweets_clean.csv'` for all tweets they assigned to a terms (e.g., “race”), then searched this subset of tweets for strings related to non-DEI uses of that term (e.g., “5K”, “nascar”, “sailing”). To ensure that the resulting tweets were not related to DEI, I eliminated any that included the entire suite of DEI-terms with which Goad and Chartwell conducted their searches (e.g., “racism”, “equality”, “gender”, “social justice”, “blm”, “equity”, see `"validation code"` in (*5*)). Note that this method provides a conservative estimate of any non-DEI tweets that were included in Goad and Chartwell’s analyses, as it will only capture tweets using the non-DEI terms for which I searched. The complete list of filtering strings for each of the 7 DEI-terms I reviewed can be found in `'twitter_errors.R'`; the file `'twitter_notdei.csv'` contains the collection of non-DEI tweets returned by this algorithm (See Table 1 for examples) is.

The seven search terms reviewed comprise N = 97337 tweets, which is 64.34% of Goad and Chartwell’s twitter dataset. With the conservative validation method described above, I found that 11.37% of the tweets for the seven focal terms were not actually DEI-related, with the percentage of irrelevant tweets for a given term ranging from 1.91 - 36.48% (Table 2). If there were no additional errors in these or the remaining 14 terms, the overall error rate for the entire data set would be 7.31%.

NIH and NSF grants

I found two major sources of error in Goad and Chartwell’s NSF and NIH data. First, their sample sizes for the number of grants were vastly inflated because they failed to correct for the mechanism by which these agencies transfer funds to the different institutions collaborating on a successful proposal. When a grant proposal that includes collaborators at different institutions is selected for funding, the agency will transfer each researcher’s portion of the grant’s budget directly to each institution.A single successful grant proposal may therefore be represented in the agency’s database by multiple “awards”. By not consolidating

different awards for the same proposal in their dataset, Goad and Chartwell have inflated their estimates of the number of NSF and NIH grants in their dataset by 20.55% and 200%, respectively. The file `'grants_dupes.csv'` contains all duplicated grant records.

Goad and Chartwell also failed to screen for alternative uses of their focal terms when reviewing the NSF and NIH grants. For example, N = 2936 of the NSF grants they identify as being DEI-focused when searching with the term "diversity" are actually grants for ecological or evolutionary research on genetic, phylogenetic, or species diversity (see `'grants_nsf_diversity.csv'`, see Table 3 for examples).

Scientific publications in Google Scholar

Finally, Goad and Chartwell sought to identify DEI-related publications in the scientific literature. To do so they searched the repositories Google Scholar, arXiv, Web of Science, and PubMed 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"). I reviewed their data from Google Scholar and Pubmed.

Goad and Chartwell once again failed to search their results for duplicate records. The 20537 duplicates that remained in these datasets inflated their estimate of DEI-related publications in Google Scholar and PubMed by 18.74% and 26.7%. They also failed to exclude hundreds of articles that were published in cultural studies, humanities, and legal journals (Table 4 and), as well as thousands of non-DEI articles on topics ranging from palliative care for cancer patients to transcatheter aortic valve replacements (see Table 5, and `'gs_neurology_examples.csv'`, `'pm_nondei_examples.csv'`.

Conclusion

A review of the data used in Goad and Chartwell's NAS report finds it includes thousands of duplications and irrelevant records. It is important to emphasize that the error estimates presented are conservative, as the procedures described here are merely a "first

pass" using relatively simple methods; more robust validation efforts, for example using keyword co-associations, will almost certainly identify additional errors.

Bibliography

1. M. Goad, The Ideological Intensification of DEI in STEM (2023), (available at https://www.realityslaststand.com/p/the-ideological-intensification-of).

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

3. 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).

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

5. E. M. Bruna, Code for identifying errors in datasets used in the report "Quantitative Study of Diversity, Equity and Inclusion in STEM Subjects in American Universities" (National Association of Scholars, 2022) (2023), (available at https://github.com/BrunaLab/quantdei_nas).

6. M. Goad, B. R. Chartwell, "Ideological intensification: A quantitative study of diversity, equity, and inclusion in STEM subjects at American universities." (National Association of Scholars, New York, 2022), p. 50.