Supplementary Materials for

Fundamental errors of data collection & validation undermine claims of 'Ideological Intensification' in STEM

Emilio M. Bruna

Corresponding author: embruna@ufl.edu

Materials and Methods: Cleaning and Validation of NAS data

Below I present a brief overview of the methods used to review the contents of 5 datasets used by Goad and Chartwell to visualize trends in DEI-language use. These datasets can be found in the `'out/twitter'`,`'out/grants'`, and `'out/scholarship'` folders of the NAS Report's Github repository (Goad 2023).

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 datasets, I wrote code in the R statistical programming language (R Core Team 2020) to conduct some simple data validation tests. This code, which included functions from the `tidyverse` (Wickham et al. 2019), `textedit` (Rinker 2018), and `janitor` (Firke 2021) libraries for cleaning, filtering, de-duplicating, and summarizing data frames, is available on Github (https://github.com/embruna/quantdei_nas). The github repository also includes `.csv` files of output of these validations (e.g., lists of duplicated records). Below I provide summaries and representative examples of the errors revealed by the validation tests.

1. University Twitter accounts

Methods

Goad and Chartwell[1] searched 895 university accounts for over 20 terms they define as DEI-related (Goad and Chartwell 2022). 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 term (e.g., "race"), then searched this subset of tweets for strings related to non-DEI uses of that term (e.g., "5K", "nascar", "sailing", "swim", "ncaa","cross country"). 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"), along with some additional terms that review of the output could be interpreted as DEI-related[2]. Note that this method provides a conservative estimate of any non-DEI tweets that were included in Goad and Chartwell's analyses, as it only captures tweets using the non-DEI terms for which I searched. The code with the complete list of these terms can be found in `"validation code/twitter_errors.R"`, while the file `'validation_output/twitter_notdei.csv'` contains the non-DEI tweets returned by the algorithm (see also Table S1 for examples).

Results

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 11.9% 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 1.89 - 36.7% (Table 2).

[1]'Bruce R. Chartwell' is a pseudonym, see Footnote 1 on https://www.nas.org/reports/ideological-intensification/full-report

[2]Terms used to exclude potential DEI-related tweets: "1619 project", "advocacy", "ally", "justice", "privilege","diversity","diverse""anti-racism", "antiracism", "bias", "black lives", "black lives matter", "blm", "civil right", "critical race theory", "culturally sensitive", "discrimination", "equality", "equity", "gender", "george floyd", "inequality", "implicit bias", "indigenous", "inclusion", "intersectional", "inclusive", "kendi", "microaggression","minority", "multicultural", "oppression", "racism", "racial", "racist", "reform", "social justice", "social change", "systemic racism", "transgender", "underrepresented", "white fragility", "white supremacy"

2. NIH and NSF grants

Methods

A review of Goad and Chartwell's data for gathering and processing NSF and NIH data and the resulting output revealed two potential sources of error. First, 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 could vastly inflate their sample sizes for the number of DEI-related grants awarded by NSF and NIH. They also failed to verify that the grants returned by their search were in fact DEI-related.

I searched for potential duplications in the `'nsf_all_grants_summary_data.csv'` and `'nih_parsed_all.fst'` files by filtering for grants with identical titles (NSF) or title and program officer responsible (NIH). The exceptions were records for which the title provided was the name of the program making the award (e.g., Postdoctoral Fellowship program, Graduate Reserach Fellowship program, Waterman awards); all of these records were maintained. The file `'grants_dupes.csv'` (Bruna 2023) contains all duplicated grant records.

To search for the potential inclusion in their dataset of non-DEI awards, I filtered to include on NSF grants they flagged as "DEI-Diversity", and excluded all grants whose titles included the DEI-related terms applied to the Twitter dataset. I also conducted a narrower search by filtering with a set of terms frequently used in the titles of grants investigating ecological or evolutionary diversity. The resulting datasets are `'validation_output/grants_nsf_diversity_wide.csv'` and `'validation_output/grants_nsf_diversity.csv'` (Bruna 2023). Code for both of these analyses is at `"validation code/grant_errors.R"` (Bruna 2023).

Results

By failing to consolidate financial awards to collaborators working on the same grant, Goad and Chartwell inflated their sample sized by 20.55% and 200%, respectively. After deduplicating the awards from NSF and reviewing those they flag as DEI-related, I found that at least N = 1882, and possibly as many as 7046 of these are actually grants for ecological or evolutionary research on genetic, phylogenetic, or species diversity (see Table S3 for examples). This represents 25.42-95.18% of the grants in this DEI category. This represents 25.42-95.18% of the grants in this DEI category.

3. Scientific publications in Google Scholar

Methods

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 (`'google_scholar.fst'`) and Pubmed (`'google_scholar.fst'`) for duplicates and to verify the journal titles using procedures similar to those for Twitter and grant data (see `"validation code/publication_errors.R"` and output files `'gs_neurology_examples.csv'` and `'pm_nondei_examples.csv'`).

Results

Goad and Chartwell once again failed to search their results for duplicate records. As a result the 2.0537×10^4 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 S5).

4. Conclusion

The data used in Goad and Chartwell's NAS report 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

Bruna EM. 2023. Code for identifying errors in the NAS report "Quantitative study of diversity, equity and inclusion in STEM subjects in American universities" (https://github.com/BrunaLab/quantdei_nas).

Firke S. 2021. Janitor: Simple tools for examining and cleaning dirty data.

Goad M. 2023. The Ideological Intensification of DEI in STEM.

Goad M, Chartwell BR. 2022. Ideological intensification: A quantitative study of diversity, equity, and inclusion in STEM subjects at American universities. National Association of Scholars (https://www.nas.org/reports/ideological-intensification/full-report).

R Core Team. 2020. R: A language and environment for statistical computing.

Rinker TW. 2018. Textclean: Text cleaning tools.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. 2019. Welcome to the Tidyverse. Journal of Open Source Software 4: 1686.

Table S1: Minimum number of irrelevant tweets attributed by Goad and Chartwell to seven different DEI terms, the total number of tweets for each term in their original dataset, and the minimum percentage of irrelevant tweets.

DEI Term	**Irrelevant Tweets (N)**	**Total Tweets (N)**	**% Irrelevant**
diversity	502	26499	1.89
equity	454	11883	3.82
justice	2270	21707	10.46
advocacy	729	6311	11.55
race	5763	25187	22.88
ally	515	2074	24.83
privilege	1349	3676	36.70

Table S2: Sample tweets erroneously considered DEI-related twitter activity. Tweets have been truncated at 140 characters and all twitter handles, urls, emojis, and emoticons have been removed for clarity; complete tweets are at (5).

Term	**Ex**	**Tweet**
advocacy	1	a passionate physician and educator commited ot medical education, patient advocacy and community medicine, @– awards sarah coles ' ' its alumna of the year honor.
	2	rsvp today for a day of legislative advocacy at the state capitol! fundazstudents
	3	join fellow wildcats for a day of legislative advocacy at the arizona state capitol on march . fundazstudents
	4	work with uofa state relations to advocate for the university at the state legislature. become an advocat:
	5	the basic trial advocacy class at the @– school argued their case in a mock trial on saturday, nov. .
ally	1	@– y grades can be given for students if the faculty member decides and student approves in writing. please visit for details (look under the students-academic section)
	2	@– congratulations!
	3	@– welcome to the sun devil family!
	4	@– @– congratulations!
	5	@– @– @– congrats on your accomplishments, asugrad! we're proud to celebrate with you today!
diversity	1	asuyearinreview: arizona has the greatest diversity of rattlesnakes anywhere in the world. @– snake e tongue sticking out ert dale denardo offers these tips on what to do and not do if you encounter one.
	2	rt @–: mya, an asteroid hit the yucatn, killing ~% of earth's life diversity. tonight learn the stories it left in the ground.
	3	@– there are , known species of ants. actual number is probably x that. diversity of their social organization is remarkable
	4	a new university of arizona-led study uses big data to assess why the diversity of species varies across the globe. hint: it's not just about temperature. what researchers learned changes our understanding of future diversity in a warming world.
	5	a new study co-authored by university of arizona researchers provides the first quantitative assessment of how environmental policies on deforestation, along with forest fires and drought, have impacted the diversity of plants and animals in the amazon.
equity	1	access to clean water should be a human right, but there is a price for providing it. the @–'s kyl center for water policy recently released "" tenets of water equity,"" discussing this very issue.
	2	rt @–: haven't seen our retirement and personfinance blog @– @–? please check it out! recent articl
	3	""highly speculative:"" prof. renee jones talks to @– about private equity ""unicorn"" start ups and the dangers of deregulation
	4	rt @–: congrats!! alex mancebo, @– boston office, focuses his practice on private equity, m a, and other complex business tran
	5	rt @–: thanks to benjamin clinger of @– for his crash course on private equity m a at today's @– lea

justice	1	@– will honor the legacy of supreme court justice sandra day oconnor with the national premiere of sandra day oconnor: the first. you won't want to miss this special documentary!
	2	.@– researchers have found that there is a higher likelihood of receiving a false guilty plea during the covid pandemic. read more about how the criminal justice system has changed during the pandemic
	3	two weeks before her first year at asu, carson swisher changed her major, and it changed her life. now the asugrad has a criminal justice degree from @– and hopes to work in the legal system as a prosecutor and then a judge!
	4	.@–'s home in washington, d.c., is the first building in the nation's capital named for two remarkable women: retired supreme court justice sandra day o'connor and former u.s. secretary of the air force barbara barrett.
	5	.@–'s popular bachelor's program in justice studies is now available through @–, creating additional opportunities for students to pursue a degree. : file
privilege	1	rt @–: years ago today, beardown was born. it is a privilege to recognize the legacy of john byrd ""button"" salmon beardown f
	2	rt @–: thank u @– @– for this very special honor. it's a privilege to work with all of you @– @–
	3	rt @–: two of the greatest guys i have ever had the privilege of working with over the years. great represenatives of @–
	4	rt @–: i had the privilege of popping up on kids a couple days ago.. blessed
	5	rt @–: years ago today, beardown was born. it is a privilege to recognize the legacy of john byrd ""button"" salmon beardown f
race	1	join the @– for the jeff coombs memorial virtual road race and boston marathon celebration.
	2	ronald a. wilson, ua title ix director and a former presiding judge for the city of south tucson, will speak about the historical relationships between the law and race in the u.s. on feb. , - p.m. the lecture is free and open to the public.
	3	join the @– in the jeff coombs memorial road race on sept. . register here:
	4	rt @–: artificialintelligence wont be spawning supercomputers or robots programmed to end the human race. ai will be working with us
	5	good luck to former uofa student and @– champ @– as he attempts to race in both the indy and coke . beardown!

Table S3: Ex non-DEI NSF grants that were included in the NAS database as 'DEI: Diversity-related'.

Ex	Grant Title
1	estimation & observation of stochastic biochemical networks
2	workshop proposal for deep time earth-life observatories (detelos)
3	applying bathymetric lidar to advance marine landscape ecology in the third dimension
4	achieving heightened goals: undergraduate research in ecology at the mountain research station
5	integrative biology and ecology of marine organisms
6	vision 2020: an open space technology workshop on the future of earthquake engineering; st. louis, missouri; january 2010
7	summer fellowships in biogeochemistry and climate change
8	network for earthquake engineering simulation - reducing seismic vulnerability
9	undergraduate research experiences in tropical conservation science
10	the cepob3b young cluster: a new laboratory for studying the role of environment in planet formation and cluster evolution
11	diversification and evolution of major trophic modes in the xylariaceae: exploring the role of previously unknown symbiotrophic and saprotrophic fungi
12	plant use and domestic economy among eurasian mobile pastoralists: semirech'ye, kazakhstan during the bronze and iron age interface
13	plant-herbivore community assembly and the problem of specificity: do insect herbivores specialize among sympatric, congeneric plants in tropical forests?
14	factors that influence the amount and pattern of genetic diversity in zymv
15	the consequences of global events on vertebrate biodiversity: the paleozoic actinopterygian radiation
16	the latitudinal gradient in plant diversity: evidence from the sedges.
17	integrating morphology, molecules and ecology to understand diversification and species coexistence within the madagascar olive, noronhia (oleaceae)
18	characterization of foliar fungal endophyte communities of sequoia sempervirens and investigation of their symbiotic relationship
19	plant chemical defenses and nectar traits mediating floral competition
20	hydrological controls of riverine ecosystems of the napo river (amazon basin): implications for the management and conservation of biodiversity

Table S4: A sample of non-STEM journals with articles that were treated as DEI-publications in STEM outlets (with the number of articles from each).

Repository	**Source**	**N**
Google Scholar	race ethnicity and education	48
	race & class	25
	science education	25
	educational studies in mathematics	24
	journal of chemical education	19
	cbelife sciences education	17
	physics teacher	17
	educational researcher	14
	cultural studies of science education	13
	physical review physics education research	13
	annual review of law and social science	12
	journal of mathematics teacher education	12
	race, gender & class	12
	teachers college record	12
	teaching race and anti-racism in contemporary	12
	urban education	12
	cambridge journal of education	11
	critical sociology	11
	journal for research in mathematics education	11
	journal of negro education	11
PubMed	j law med ethics	142
	int j law psychiatry	122
	j urban health	108
	j health polit policy law	92
	hosp law newsl	89
	law hum behav	86
	behav sci law	80
	j am acad psychiatry law	66
	med law	64
	j law med	49
	contraception	45
	am j law med	40
	j contemp health law policy	38
	health law vigil	33
	annu rev popul law	31
	med sci law	31
	j health hosp law	30
	med law rev	30
	law med health care	27
	aids policy law	23

Table S5: Sample non-DEI articles included by Goad and Chartwell in their analsyis of DEI-publications in STEM journals.

Repository	**Ex**	**Title**	**Year**	**Source**
Google Scholar	1	translating the biology of aging into novel therapeutics for alzheimer disease	2019	neurology
	2	revisiting protein aggregation as pathogenic in sporadic parkinson and alzheimer diseases	2019	neurology
	3	revised airlie house consensus guidelines for design and implementation of als clinical trials	2019	neurology
	4	novel biomarker signatures for idiopathic rem sleep behavior disorder a proteomic and system biology approach	2018	neurology
	5	the biology of cutaneous neurofibromas consensus recommendations for setting research priorities	2018	neurology
	6	serum neurofilament light in familial alzheimer disease a marker of early neurodegeneration	2017	neurology
	7	the autism epidemic ethical legal and social issues in a developmental spectrum disorder	2017	neurology
	8	biological tumor volume in 18fetpet before radiochemotherapy correlates with survival in gbm	2015	neurology
	9	dystrophin quantification biological and translational research implications	2014	neurology
	10	defining the clinical course of multiple sclerosis the 2013 revisions	2014	neurology
PubMed	1	exploring us shifts in antiasian sentiment with the emergence of covid19	2020	int j environ res public health
	2	a critical review of theory in breast cancer screening promotion across cultures	2008	annu rev public health
	3	navigating uncertainty employment and womens safety during covid19 reflections of sexual assault resistance educators	2020	gend work organ
	4	chronographic theory of development aging and origin of cancer role of chronomeres and printomeres	2015	curr aging sci
	5	there is a balm in gilead black social workers spiritual counterstory on the covid19 crisis	2020	soc work public health
	6	like i have no choice a qualitative exploration of hiv diagnosis and medical care experiences while incarcerated and their effects	2019	behav med
	7	interventions that retain african americans in hivaids treatment implications for social work practice and research	2015	soc work
	8	hivaids a minority health issue	2005	med clin north am
	9	lower hiv prevalence among asianpacific islander men who have sex with men a critical review for possible reasons	2011	aids behav
	10	culture in cancer survivorship interventions for asian americans a systematic review and critical analyses	2021	asian am j psychol