Diversity in Scientific Teams Researching COVID-19

Richard B. Freeman NIH May 17, 2022,
based on work with Qingnan Xie, Sifan Zhou and Sen Chai

Definition: Diversity in team is nearness of attributes to **random choice** from **relevant population** by a measure of difference of attributes or of 1 – concentration. Regression analyses often relate diversity to observables.

Alternatively, we can measure diversity of team as **presence** of at least one researcher from an identifiable community on the notion that one person suffices to connect the community and the research.

Least diverse: single authored paper. But individuals can differ in diversity. Author has 2 addresses; master's in math and PhD in sociology in 3 countries is more diverse than author one address, field, and country experience.

Most diverse: unlikely to be optimal. People with special attributes may be necessary to create team, such as diaspora/returnee on country collaboration.

Output question: What is optimal diversity in inputs for outcomes? Diversity can raise social value independent of bibliometric measures.

Diversity Attributes and Metrics

Diversity of people inputs on a paper/patent/innovation by:

- Address
- Gender
- Academic age
- Name ethnicity/place of birth
- Discipline/topic in keywords/reference
- Education/other background

Outcome "indicators" of impact of diversity – citations, IF of journal, prizes. Big confounder for citation – "self-citing" network bias: Milard (2014) found that 75% references were to people authors knew. Would your paper gain more cites with co-author from Iceland or China?

Our approach: Consider attributes/indicators "appropriate" to question. Eager to hear other ways to develop metrics and use data to illuminate policies, issues, or to exploit "natural experiments".

Main weakness of analysis is that we take team as given when people endogenously form teams → selection bias issues. One test of team is its life cycle. At minimum length of collaboration reflects quality based on the judgement of the authors.

1) International Diversity in COVID-19 Vaccine Development

- **Advanced COVID-19 Vaccines/Candidates and Developers**

Country	*Developer*	*Vaccine Type*	*Approvals in #countries*	*Clinical Trials*
US	Moderna + NIAID	RNA	85	46 trials in 20 countries
US	Novavax	Protein Subunit	36	11 trials in 7 countries
US/Germany + China	Pfizer/BioNTech + Fosun Pharma	RNA	137	61 trials in 24 countries
Belgium (US)	Janssen (Johnson & Johnson)	Non Replicating Viral Vector	106	19 trials in 18 countries
UK	Oxford/AstraZeneca	Non Replicating Viral Vector	138	58 trials in 30 countries
China	Clover	Protein Based	Under trial	Phase 3

a) Most firm leaders have international backgrounds

Company	**#C-suites**	**%C-suites with**	
		International education	**Any non-English/German name**
Moderna	15	60% (non-US)	73% (non-English)
Pfizer	13	54% (non-US)	62% (non-English)
BioNTech	6	67% (non-German)	83% (non-German)

Moderna co-founder *Derrick Rossi* born in Toronto in a Maltese immigrant family, Ph.D. from the University of Helsinki. In 2003, post-doc at Stanford University, 2003-7. Associate Prof Harvard Medical School. 4 of 6 non-English name C-suites of Moderna have US education

Pfizer CEO *Albert Bourla* born in Thessaloniki, Greece, in family of Sephardic Jews. Greek PhD, promoted by Pfizer to US, immigrant in 2001.

BioNTech co-founder

b) LinkedIn & name based international background and gender of inventors on vaccine patents

Company (#patents; #inventors)	%Patents with (base on LinkedIn)			%Inventors on Patents with			
	At least one inventor with international work or education	At least one inventors' name ethnicity different from the major name ethnicity of address country	At least one female inventor	At least one international work or education	Name ethnicity differs from the address country	Union of work, education and ethnicity	Female name
Moderna (11; 29)	91%	100%	55%	59%	66%	68%	24%
Pfizer (9; 36)	78%	100%	78%	58%	50%	77%	29%
BioNTech (20; 59)	95%	100%	90%	66%	45%	71%	47%
Total (40;124)	90%	100%	78%	62%	51%	72%	36%

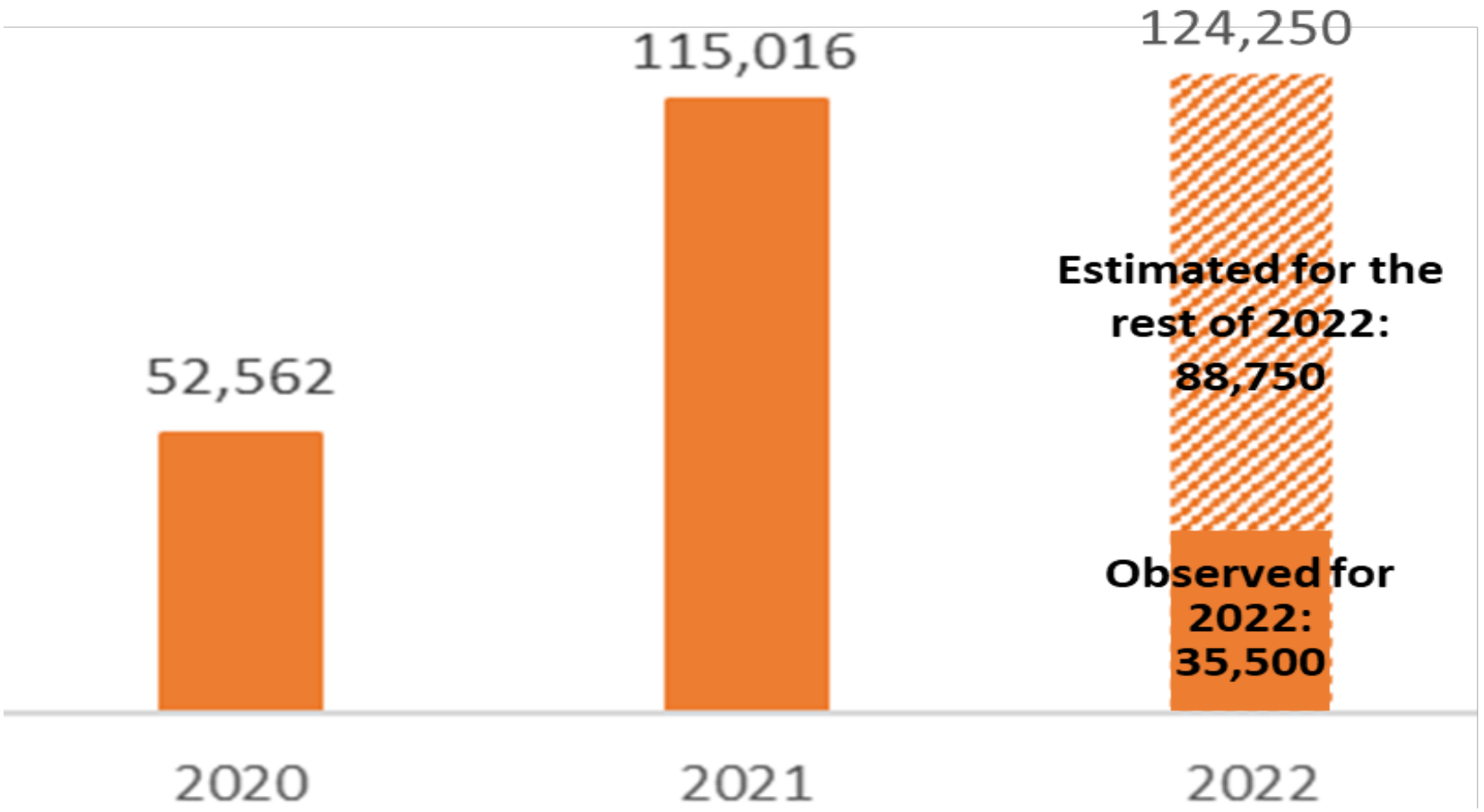
Note: 20 BioNTech joint TRON Patents with 43 inventors excluded due to a lack of available LinkedIn data on TRON scientists.

c) Address & name based international background and gender of authors on clinical trial papers

Company (#trial papers; #author per paper)	%Papers with				%Authors on papers with			
	Multi-national addresses	At least one author with papers published with 2 or more country addresses	At least one authors' name ethnicity different from the major name ethnicity of address country	At least one female author	Papers published with 2 or more country addresses	Name ethnicity different from name ethnicity on address country	Union of address and ethnicity	Female name
Moderna (*8; 29*)	13%	100%	100%	100%	36%	49%	70%	34%
Novavax (*6; 81*)	50%	100%	100%	100%	20%	22%	36%	51%
Pfizer/BioNTech (*9; 55*)	78%	100%	100%	100%	31%	33%	50%	34%
J&J (*3; 36*)	100%	100%	100%	100%	58%	40%	71%	36%
Oxford/AstraZeneca (*19; 337*)	68%	100%	100%	100%	18%	19%	32%	55%

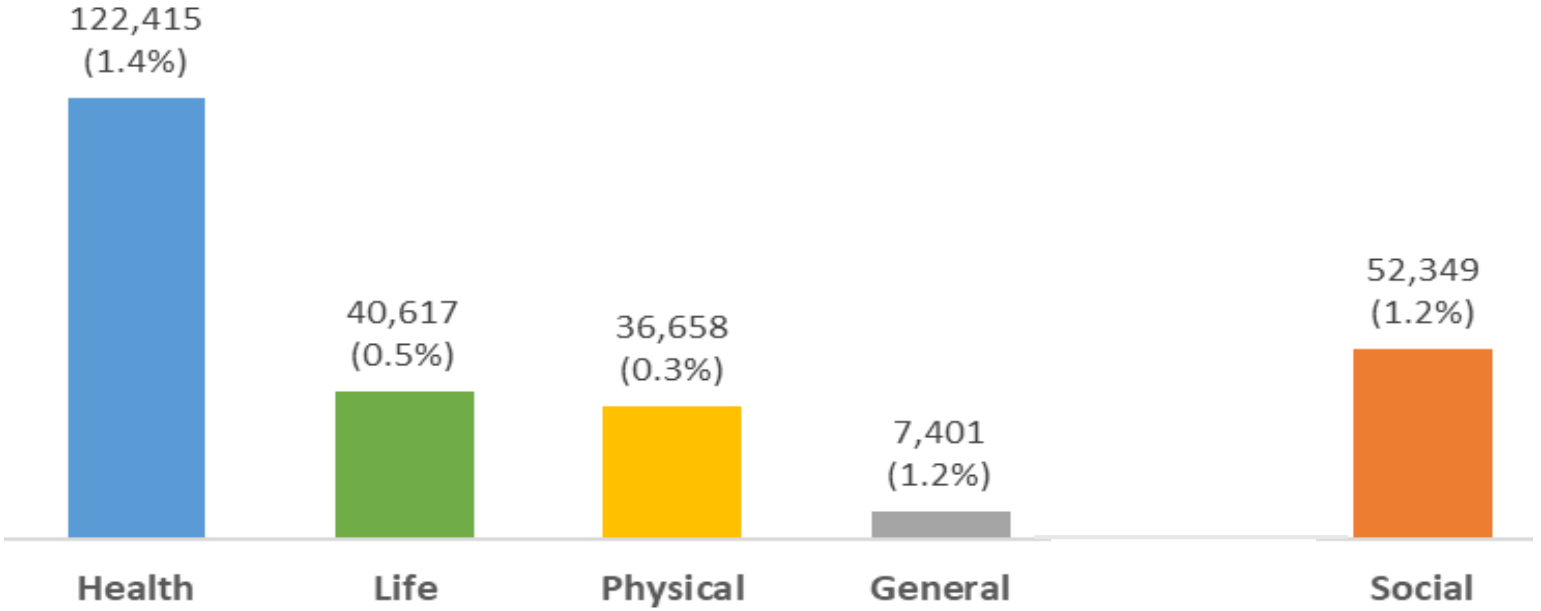
2) Diversity of Explosion of COVID-19 journal articles

a) #COVID-19 articles in 2020-2022 (vs 9,903 coronavirus papers from 2000-2019)

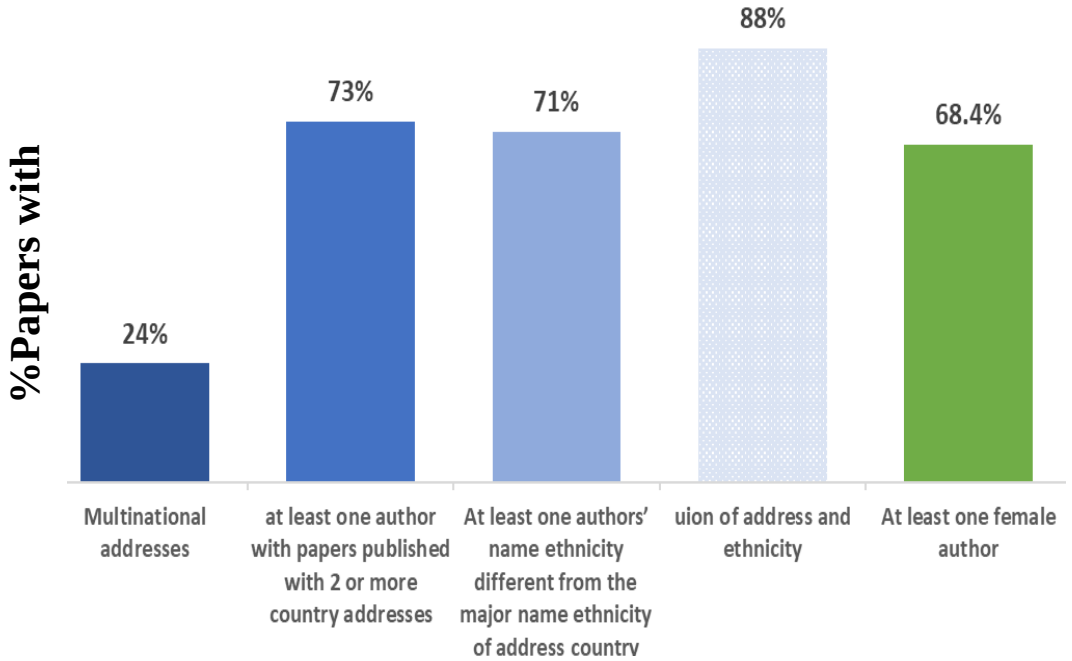


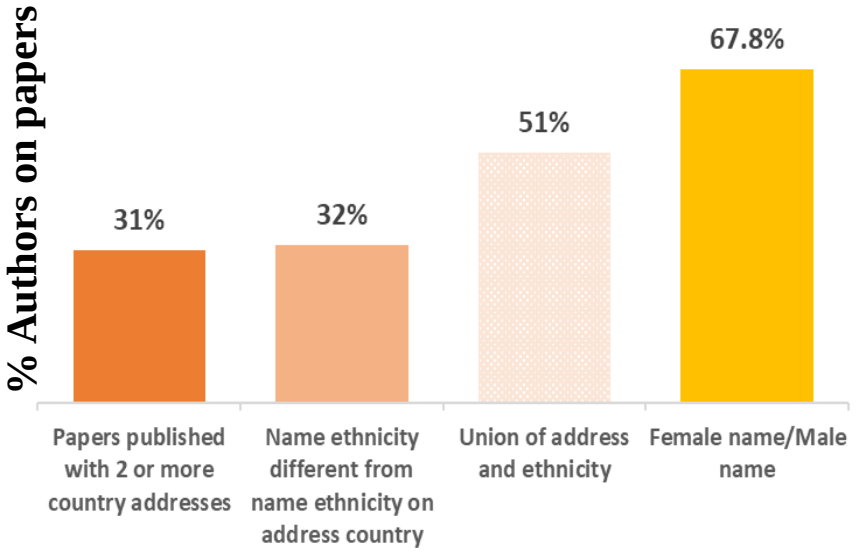
Note: Based on Scopus data We found 35,500 COVID-19 journal articles in the mid April 2022. We estimate the number of COVID-19 papers in 2022 by multiplying 35,550 with 3.4 (105 days through the mid April/365 days of the year). Because some papers are put online earlier than their publication date, there should be less new 2022 papers indexed into Scopus later this year →over estimate the 2022 number.

b) Field Diversity: # field tags COVID-19 journal articles (COVID-19 percentage of all tags in field)



c) Address & name based international background and gender of authors on COVID-19 papers in Natural Science fields





Note: There are 15,699 authors on 2,000 randomly sampled COVID-19 papers in Natural Science fields, and 10,775 (69%) with valid gender predictions. One reason for no gender prediction could be the program cannot tell genders for some certain name (like Chinese name or name could be both female and male, and the other is lacking of valid first name info from Scopus.)

Regression coefficients (std errors) of relation between diversity in papers and CiteScore of publication journal and Citations

Dependent Variable (# observations)	**Ln(CiteScore)** (1,864)	**Ln(citation)** (1,126)
# of Country addresses	-0.028 (0.019)	0.001 (0.03)
# of Authors on papers	0.023 (0.002)	0.020 (0.003)
Proportion of authors with name ethnicity different from major name ethnicity of address country	**0.167 (0.066)**	**0.071 (0.115)**
Proportion of authors with earlier papers at different country address	**0.505 (0.078)**	**0.654 (0.133)**
Proportion of authors with female first name	**0.038 (0.091)**	**-0.115 (0.16)**
Proportion of authors with unidentified gender or first name	-0.098 (0.077)	0.246 (0.133)
Open Access journal	0.500 (0.06)	0.483 (0.123)
English language paper	1.724 (0.098)	0.93 (0.201)
20 Field Dummy variables	yes	yes
Year	yes	yes
Cons	-1.052 (0.135)	-0.946 (0.29)
Adjusted R2	0.3648	0.2547

Note: Based on 2,000 randomly sample of 148,296 COVID-19 papers in Natural Science fields.

3) Gender Citation Homophily (with Sifan Zhou and Sen Chai)

CAREER NEWS | 29 July 2021

Fewer citations for female authors of medical research

Papers by women in elite medical journals are half as likely to be cited as are similar articles authored by men, research finds.

Our data: 3 year forward citations received by 2,432,806 US-based English language journal papers published between 2002 and 2017 retrieved in PubMed, then matched with Microsoft Academic Graph. Gender of papers set by first or last author.

Figure 1B. Forward citations received, focal article gender is classified by *last* author

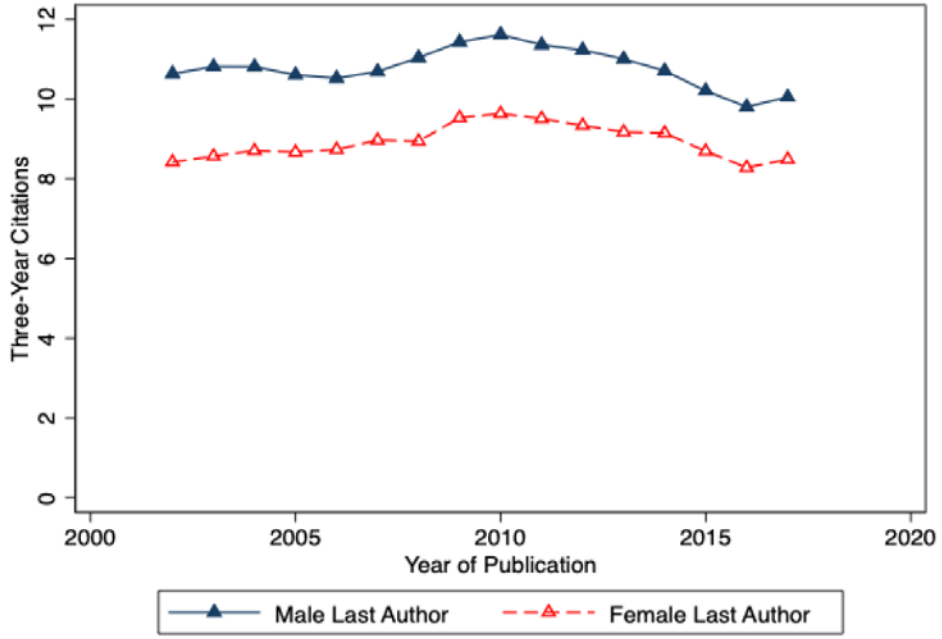
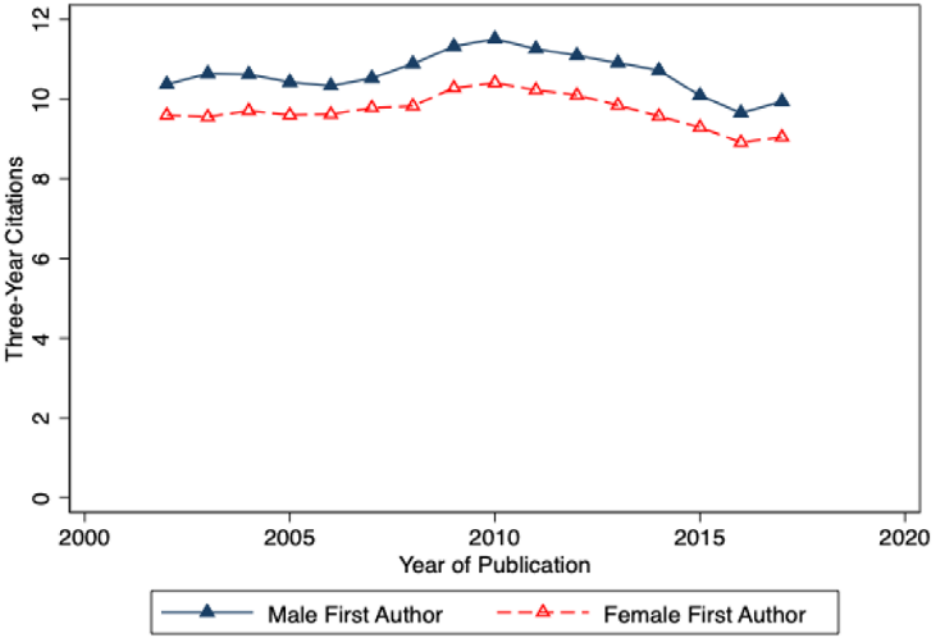


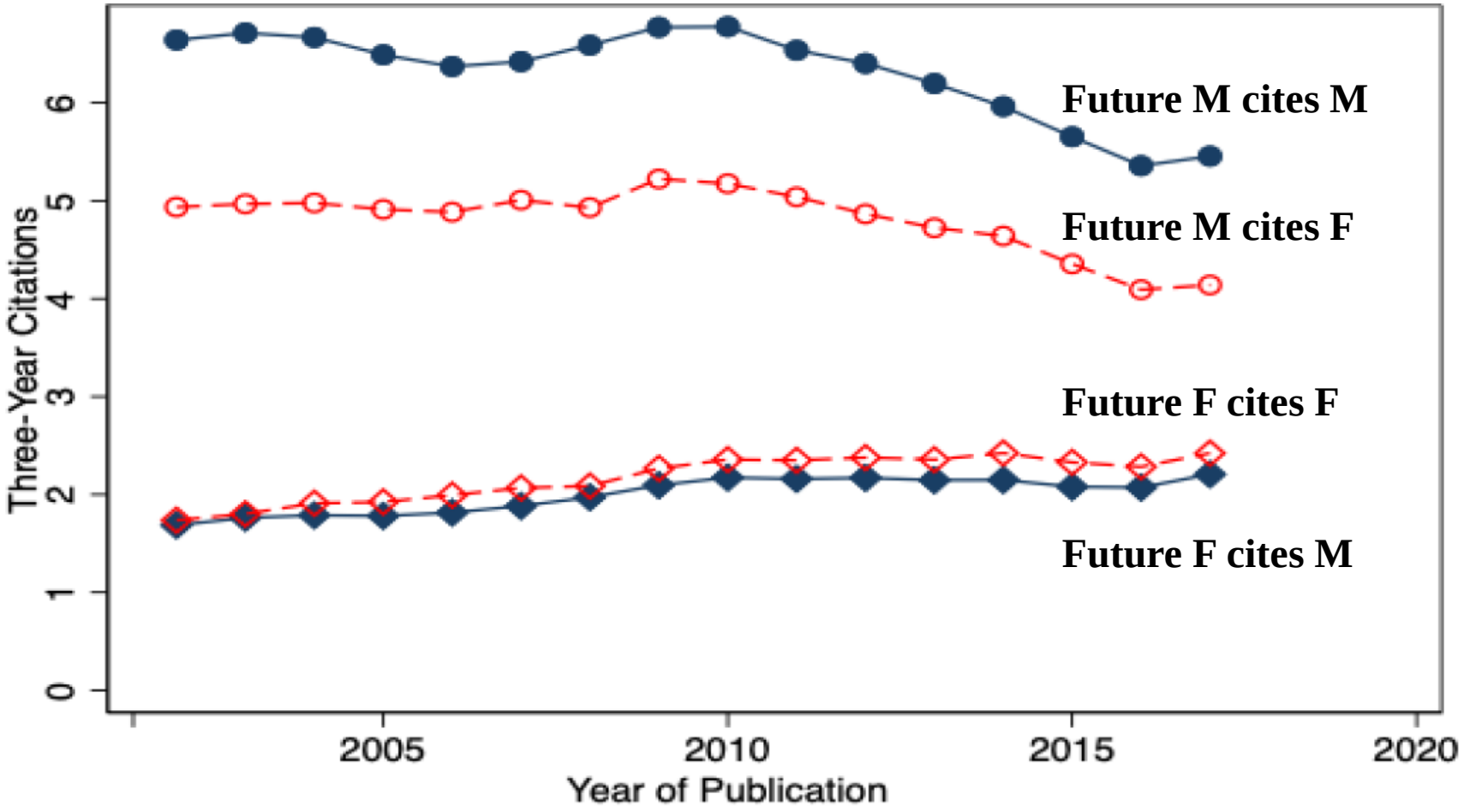
Figure 1C. Forward citations received, focal article gender is classified by *first* author



Gender Homophily in Citation: Gender-led articles more likely to get future cites from same gender-led articles

Blue: cited paper is M; Orange: cited paper is female

Citing and cited papers gender by last author



Note: When we looked the citing and cited papers gender by first author, the gap between the gender citations to F and M papers are smaller. Possibly because the increasing of female first author paper

How Gender Citation Homophily Hurts Minority Gender: "minority scale bias"

Groups M and F write same "quality" papers which have same chance of **being cited** absent group tendency to cite from own group. Homophily leads each to cite own papers by "quality" + h/2 and disfavor the other group by –h/2. With % F of papers the M-F difference in citations will be **h (1- 2 %F)**, so homophily bias balances out if %F = 1/2. But if %F < 1/2 , F papers will receive fewer citations than same quality M papers.

Minority scale bias (MSB) is the bias in citations due to citation homophily among groups differing in number of papers.

If F and M have different own group preferences, denoted as ½ h_m and ½ h_f, difference in citations will also depend on the difference between own group preferences, $h_m - (h_m + h_f)$ **%F.** If minority group had greater homophily could overcome MSB. If majority group had greater homophily MSB would be bigger.

Conclusion and Further Work

Diversity substantial in science along many dimensions. Homophily also present as scientists form teams and cite those they know.

"Optimal diversity" difficult to determine, likely differs with topic, team, type of diversity, networks of who knows who based on other factors. Breadth of maximand.

Minority scale bias in citations harms minorities, such as women and researchers from small countries, with likely impacts on their careers.

Diversity along some dimensions associated with higher CiteScores/ citations; but homophily along some dimensions also likely to have positive effect as in specialists in particular area.

Future work: Measure diversity of background/ideas and knowledge of co-authors. Analyze endogeneity of teams and team lifecycle – number/length of collaborations. Estimate effect of gender citation homophily on careers. Seek ways to encourage diversity to reduce undesirable homophily effects → optimum.