SUPPLEMENTARY MATERIALS FOR

GENDER DIFFERENCES IN NSF OCEAN SCIENCES AWARDS

By Ivan D. Lima and Jennie E. Rheuban

PART 1. DATA COLLECTION AND PREPARATION

The NSF-OCE award data were downloaded directly from the NSF website (https://www.nsf.gov/) using the site's Advanced Search feature. Abstracts and other award information tend to be missing for records prior to 1987, so we focused on awards from 1987 to 2019. Prior to doing any processing, we removed duplicate records and cleaned the data set to eliminate typos, misspellings, inconsistencies (e.g., upper/lower case, abbreviations, dashes) and extraneous information (e.g., award numbers, PI names, HTML tags) from the awards' titles and abstracts. This was necessary to correctly group records that were part of the same Collaborative Research project, to reduce noise in the vocabulary, and improve topic model performance. In Collaborative Research awards, PIs from different institutions work together on the same research project. Each participating institution receives a separate award, but they are all part of the same project and therefore have the same title and abstract. For this type of award, we grouped the related records (by abstract) into one award, summed the amounts awarded to each organization, and combined the PI and co-PI names from each institution into one list. The NSF data did not include any information that allowed us to identify the lead PI for Collaborative Research awards. However, we assumed that the PI from the institution that received the largest amount would most likely be the lead PI for the project, and that is the criterion we used to designate the main PI for Collaborative Research awards. After grouping related Collaborative Research records, our data set contained 11,513 awards. The amount of money awarded for each project was adjusted for inflation to 2019 US dollars using the US Bureau of Labor Statistics Consumer Price Index annual average.

We inferred the gender of the PI and co-PIs for each award from their first names using a database of names and gender created with data from the following sources:

1. US Census Bureau
   https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html
2. US Social Security Administration
   https://www.ssa.gov/oact/babynames/limits.html
3. Liste de prénoms en français
   https://fr.wikipedia.org/wiki/Liste_de_pr%C3%A9noms_en_fran%C3%A7ais
4. Korean given names
   https://en.wikipedia.org/wiki/Category:Korean_given_names
5. Japanese given names
   https://en.wikipedia.org/wiki/Category:Japanese_given_names
6. Chinese given names
   https://en.wikipedia.org/wiki/Category:Chinese_given_names
7. Indian given names
   https://en.wikipedia.org/wiki/Category:Indian_given_names

For the 108 investigators whose names were not included in the database, we did a web search and identified their genders from their departmental websites. Once gender was assigned to PIs and co-PIs, we computed the percentage of women co-PIs for each award. Given that some names are gender neutral, it is possible that our gender assignment method introduces errors. However, according to the validation tests we performed, these errors should be very small in number.

NSF-OCE has different types of awards for different types of research or research-related activities. We identified the following types of awards in the data:

1. **Collaborative Research**: Larger collaborative projects involving PIs from multiple institutions.
2. **EAGER** (EArly-concept Grant for Exploratory Research): High risk-high reward exploratory research.
3. **RAPID** (Rapid Response Research): Research that is urgent with respect to the availability of or access to data, including quick-response research on natural or anthropogenic disasters and similar unanticipated events.
4. **SGER** (Small Grants for Exploratory Research): Small-scale, exploratory, and high-risk research that is urgent with respect to the availability of or access to data, including quick-response research on natural or anthropogenic disasters and similar unanticipated events.
5. **REU** (Research Experiences for Undergraduates): Support for undergraduate students to participate in active research.
6. **RUI/ROA** (Research in Undergraduate Institutions/ Research Opportunity Awards): Research support for faculty members at predominantly undergraduate institutions.
7. **CAREER** (Faculty Early Career Development): Support for early career faculty projects integrating education and research.
8. **CMG** (Collaboration in Mathematical Geosciences): Interdisciplinary, collaborative research at the intersection of mathematical sciences and geosciences.
9. **POWRE** (Professional Opportunities for Women in Research and Education): Award type established in 1997 to promote women's participation in science and engineering.

We defined "standard" awards as the typical, traditional NSF-OCE grants that are not part of the specific categories and types mentioned above. SGER is an older type of award that combines elements from EAGER and RAPID awards but is more similar to EAGER. For that reason, we grouped SGER and EAGER awards together as EAGER awards. Given the small number of CMG awards (20) and that they were awarded over a period of only eight years (2003–2010), we excluded these awards from the analysis. POWRE is an award given exclusively to women investigators. They were given between 1997 and 2000 and there are only 13 in our data set. For those reasons, they were also removed from the analysis. The remaining award types (Collaborative Research, EAGER, RAPID, REU, RUI/ROA, and CAREER) plus the "standard" awards comprise over 99% of the awards in the data set. Fifty RAPID awards are also Collaborative Research projects, but we classify them as RAPID. Therefore, the numbers and percentages for Collaborative Research awards in Figure 1 and Tables S3 and S4 are somewhat underestimated.

PART 2. ALLOCATION RATIO COMPUTATION

We think that the allocation ratio metric is best explained via an example. Suppose we have 10 women PIs and 40 men PIs in our data set and that they distribute their awards to four different programs (MGG, BO, PO, and CO) as shown in Table S1. Women allocate four awards to BO, which represents 40% of the awards from women PIs. Men also allocate four awards to BO, which represents 10% of the awards from men PIs. The ratio between the percentages of awards from women and men PIs in BO is 4, and it indicates that the proportion of biological oceanographers among women is four times larger than that of men. In other words, biological oceanographers comprise 40% of women PIs but only 10% of men PIs in this hypothetical data set.

A category (program, topic, or award type) with a low percentage of women PIs will not necessarily have an allocation ratio lower than 1, as the allocation ratio depends on the percentage of awards allocated by men to that category. If men apportion an even lower percentage of awards than women to that category, the allocation ratio will be greater than one. For example, in the period 1998–2008 (Figure 5b), the percentage of awards allocated by women PIs to RUI is very low (1.17%). However, men allocated only 0.5% of their awards to that same award type, resulting in an allocation ratio of 2.3.

TABLE S1. Hypothetical data set for women and men PIs. MGG = marine geology and geophysics. BO = biological oceanography. PO = physical oceanography. CO = chemical oceanography.

PROGRAM	WOMEN PIS		MEN PIS		ALLOCATION RATIO
	NUMBER OF AWARDS	PERCENTAGE OF AWARDS	NUMBER OF AWARDS	PERCENTAGE OF AWARDS	
MGG	3	30%	12	30%	1
BO	4	40%	4	10%	4
PO	1	10%	16	40%	0.25
CO	2	20%	8	20%	1
Total	10	100%	40	100%	

PART 3. TOPIC MODELING

We extracted 21 research topics from the awards' abstracts using a procedure very similar to that described in Lima and Rheuban (2018). The main differences are that, in this study, we rescaled the *bag-of-words* representation of our collection of abstracts using the *term frequency-inverse document frequency* (tf-idf) method and extracted the topics by applying non-negative matrix factorization (NMF) to the rescaled *bag-of-words* matrix. The tf-idf method rescales the word frequencies in the *bag-of-words* matrix based on how informative we expect them to be. It is based on the idea that words that appear often in a particular document (abstract), but not in very many documents are likely to be descriptive of the contents of that document. NMF is a matrix decomposition method similar to principal component analysis (PCA) in which the components and coefficients are greater or equal to zero. Therefore, this method is suitable for data where variables (features) are non-negative, such as matrices of word frequencies. The NMF algorithm is deterministic and therefore has advantages over probabilistic methods such as Latent Dirichlet Allocation (LDA) in terms of stability, consistency, and convergence. For more information on extracting research topics from NSF award abstracts, see Lima and Rheuban (2018).

PART 4. ADDITIONAL TABLES AND FIGURES

TABLE S2. Percentage changes in women's participation in NSF-OCE awards between 1987–1997 and 2009–2019 for PIs, co-PIs and the mean of the two. The numbers in parentheses in the first column represent the topic number.

TOPIC	PI	CO-PI	MEAN
workshop science international (8)	24.42	32.33	28.38
student reu science (17)	31.09	17.76	24.43
circulation woce atlantic (13)	17.96	23.22	20.59
seismic earthquake plate (19)	10.91	21.52	16.22
climate change record (21)	19.09	11.8	15.45
coral reef bleaching (7)	19.8	10.64	15.22
carbon organic co2 (9)	13.84	16.49	15.17
nitrogen fixation n2 (20)	12.84	13.98	13.41
shelf coastal transport (1)	9.21	14.99	12.1
iron fe phytoplankton (15)	11.11	11.34	11.23
ridge mantle melt (4)	8.4	11.14	9.77
sensor instrument measurement (12)	14.15	4.86	9.51
microbial phytoplankton cell (18)	14.4	3.45	8.93
trace element isotope (14)	7	10.44	8.72
wave internal turbulence (5)	6.16	10.24	8.2
sediment core organic (10)	8.34	7.78	8.06
population larval specie (6)	5.71	10.15	7.93
equipment scientific shipboard (2)	8.7	0	4.35
vessel ship operate (16)	–2.82	5.84	1.51
instrumentation university shared (3)	–4.39	7.28	1.44
hydrothermal vent fluid (11)	0.07	2.73	1.4

TABLE S3. Number and percentage of awards in each award type for the period 1987–2019.

AWARD TYPE	NUMBER	PERCENTAGE	CUM. PERCENTAGE
Standard	8,500	73.83	73.83
Collaborative Research	1,988	17.27	91.10
EAGER	349	3.03	94.13
REU	236	2.05	96.18
RAPID	211	1.83	98.01
CAREER	108	0.94	98.95
RUI	88	0.76	99.71

TABLE S4. Amount awarded and percentage of total amount awarded for each award type for the period 1987–2019. Amounts are adjusted for inflation to 2019 US dollars.

AWARD TYPE	AMOUNT AWARDED	PERCENTAGE	CUM. PERCENTAGE
Standard	$11,093,484,591.21	83.35	83.35
Collaborative Research	$1,924,975,456.67	14.46	97.81
REU	$102,516,226.11	0.77	98.58
CAREER	$64,394,968.02	0.48	99.06
EAGER	$42,089,704.65	0.32	99.38
RUI	$36,333,626.48	0.27	99.65
RAPID	$27,477,422.61	0.21	99.86

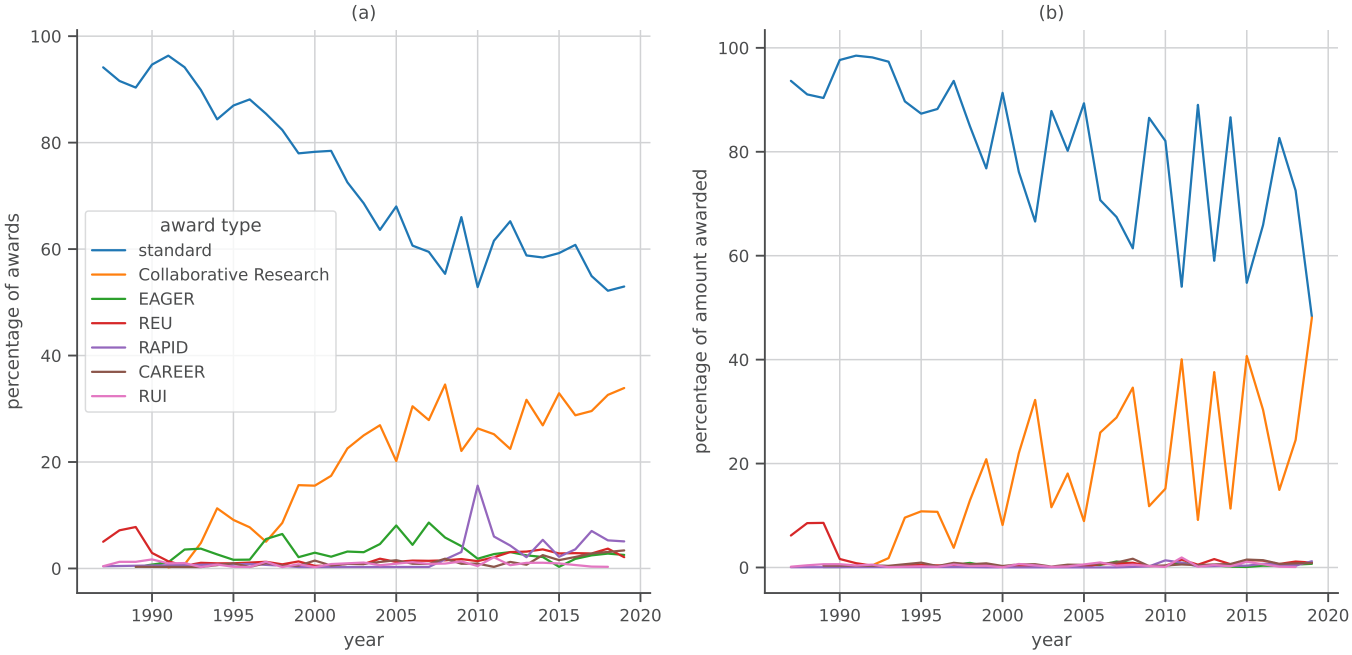

FIGURE S1. Time series of the annual percentage of the number of awards (a) and the amount awarded (b) for each award type between 1987 and 2019. The amounts awarded were adjusted for inflation to 2019 US dollars prior to computing the percentages.

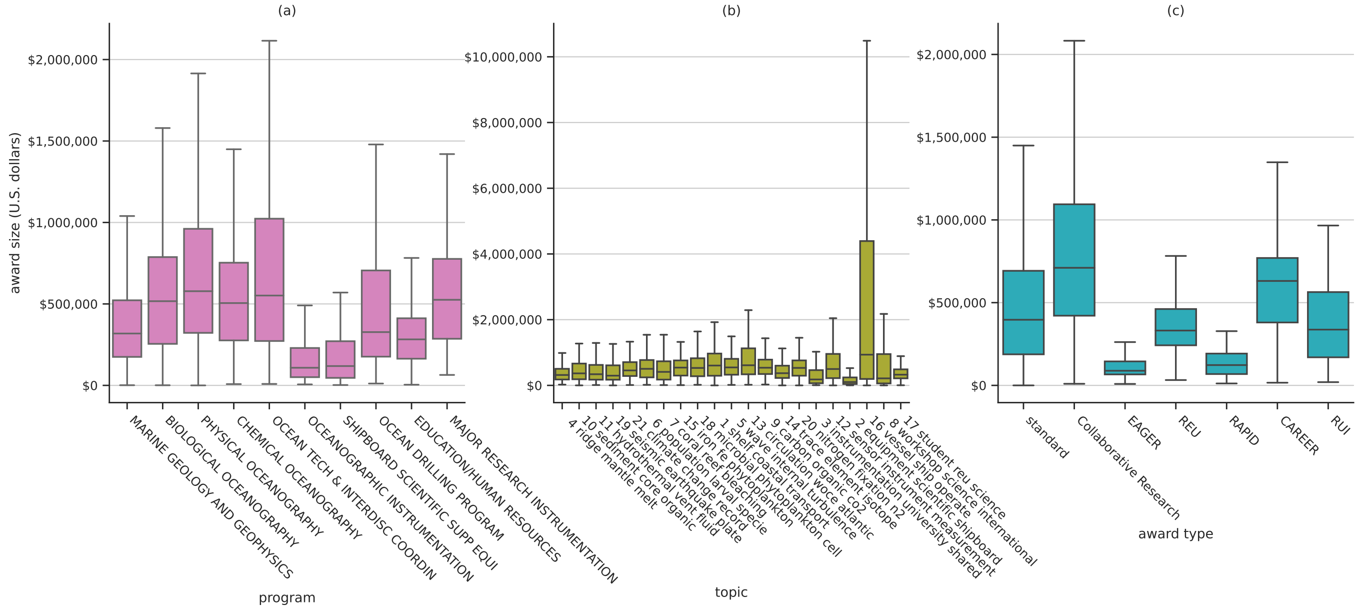

FIGURE S2. Box whisker plots of award (project) size across different (a) NSF-OCE programs, (b) research topics, and (c) award types. Amounts are adjusted for inflation to 2019 US dollars.

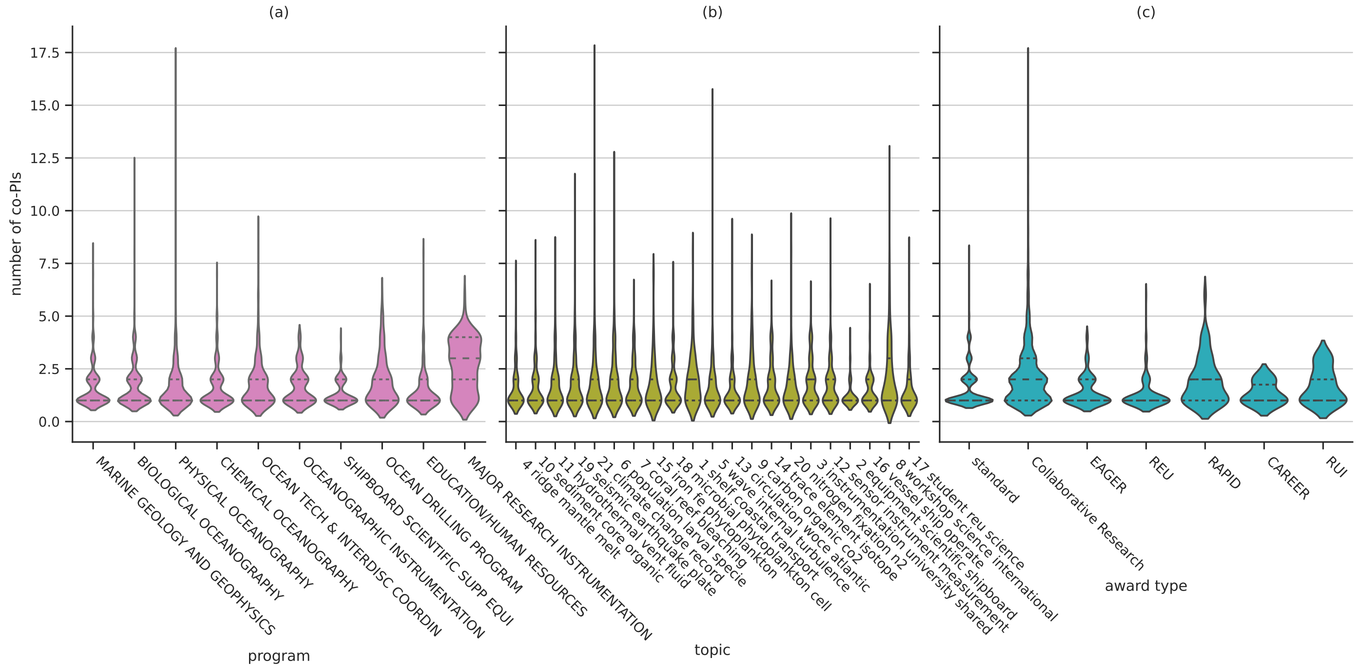

FIGURE S3. Violin plots of the number of co-PIs (team size) across (a) NSF-OCE programs, (b) research topics, and (c) award types. The dashed and dotted lines inside the violins represent the quartiles of the distribution.

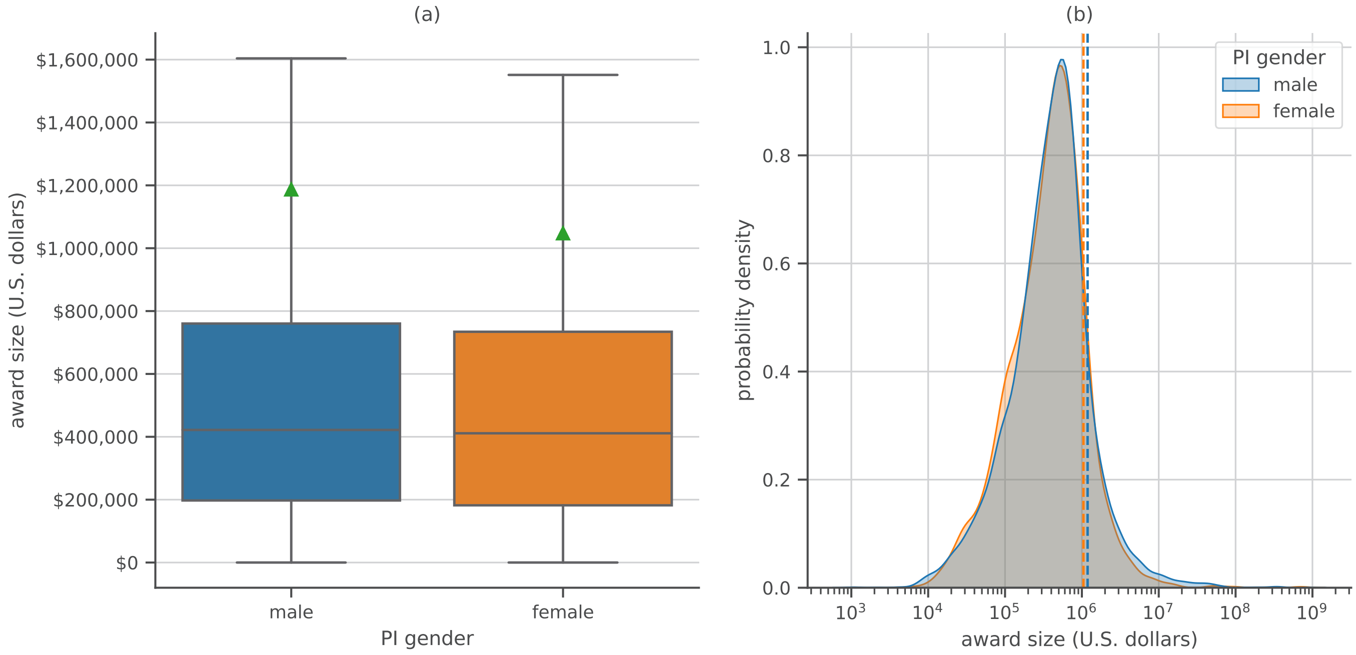

FIGURE S4. Box whisker plot (a) and probability density function (b) of the award size by PI gender. The green triangles in the box whisker plot and the vertical lines in the probability density function plot represent the means for men and women PIs. The award size (x-axis) in the probability density function plot is in a log 10 scale. A t-test shows that the difference between the means for women and men PIs is not statistically significant with a t-statistic of 0.392 and p-value of 0.695. Amounts are adjusted for inflation to 2019 US dollars.

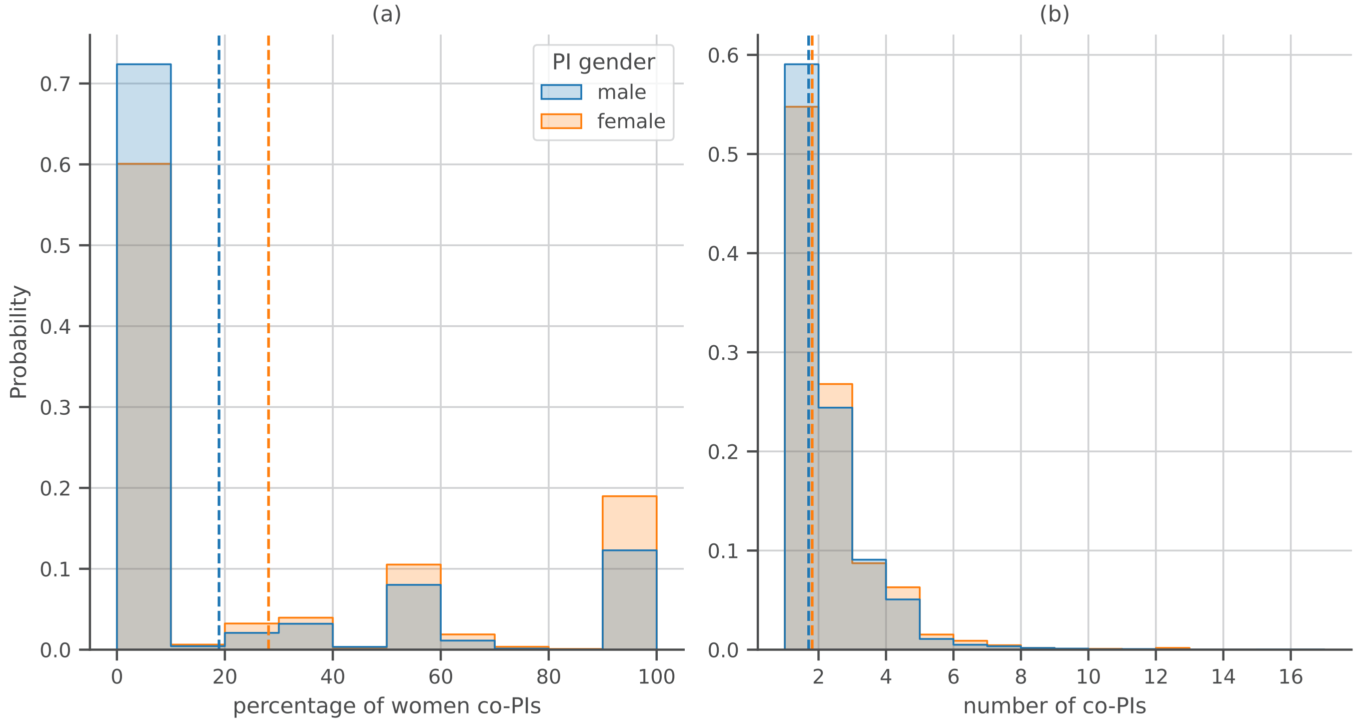

FIGURE S5. Probability distribution of the percentage of women co-PIs (a) and number of co-PIs (b) by PI gender. The vertical lines represent the means for women and men PIs. In both cases, the probability distribution for women PIs is more heavy-tailed than that of men towards higher values, resulting in a larger mean for women PIs. A t-test shows that the difference between the mean percentage of women co-PIs for women and men PIs is statistically significant with a t-statistic of 7.153 and a p-value<10^{-12}. The difference between the mean number of co-PIs for women and men PIs is small but a t-test shows that it is statistically significant with a t-statistic of 2.545 and a p-value of 0.011.