Gender, Influence, and Group Composition: A Field Experiment

Chris Karpowitz

Jessica Preece

Olga Stoddard*

January 2020

Does the gender composition of a workplace influence the ways in which team members are evaluated? Can it have long-term effects on labor market outcomes? We conducted a multi-year field experiment with a top-10 U.S. accounting program that was designed to identify the effects of work team gender composition on men and women's performance, influence, and long-term labor market outcomes. Using survey, laboratory, and administrative data, we find that women who were randomly assigned to be the lone women on a team were able to exert significantly less influence than women in majority-women teams. While their grades were not affected, women's long-term professional plans are affected by the gender composition of their group. Our findings have important implications for how workplaces approach team assignment and the evaluation of women in male-dominated settings.

*The authors may be reached at chris_karpowitz@byu.edu, jessica_preece@byu.edu, and olga.stoddard@byu.edu.
The authors thank Rachel Fisher, Dani Hogan, Connor Kreutz for outstanding research assistance. We acknowledge helpful comments and suggestions from seminar participants at University of Chicago, Norte Dame University, Brigham Young University, as well as conference attendees of the Advances in Field Experiments, Science of Diversity and Inclusion, Utah Experimental Conference, and the Western Economic Association meetings. We gratefully acknowledge funding provided by Brigham Young University and the College of Family, Home and Social Sciences.

Over the past six decades, women's labor force participation rate has risen steadily, from 34 percent in 1950 to 57 percent in 2016 (Bureau of Labor Statistics, 2017). Yet, as dramatic as these changes have been, most high-paying occupations remain largely male-dominated, and this pattern accounts for a significant portion of the gender pay gap (AAUW 2016; Altonji and Blank 1999). Not only do women remain underrepresented in high-paying jobs, but when they do enter them, they often face significant social and cultural barriers that limit their career development, undermine their authority, and isolate them (Kanter 1977; Roght 2004; Turco 2010).

Economists have increasingly studied one factor that influences women's ability to thrive in professional careers: the dynamics of workplace teams. Teams are an efficient and flexible way to coordinate production requiring a diversity of skills, talents, and information (Eckel and Grossman 2005) and have become increasingly common in the workplace (Lazear and Shaw 2007). Yet, social identity theory suggests that members of a team that is heterogeneous with respect to social categories such as gender may find it difficult to integrate their diverse backgrounds, values, and norms to work together (Jehn, Northcraft, and Neale 1999).

Indeed, scholars have found that in the lab, a male-dominated work environment leads to lower levels of female leadership because women correctly presume they will have less support from their team when men are in the majority (Born, Ranehill, and Sandberg 2018). Further, gender composition of a work network and its patterns of communication predict women's leadership (Yang, Chawla, and Uzzi 2019) and men in gender-mixed work teams disproportionately leave "non-promotable tasks" to women to complete (Babcock et al. 2017). Other lab experimental work finds that evaluations of competence are highly gendered (Isaksson 2018; Preece 2016;

Thomas-Hunt and Phillips 2004). And, the gender composition of an environment and how a task or domain are stereotyped matter a great deal for these perceptions (Bordalo et al. 2019; Coffman 2014; Karpowitz and Mendelberg 2014).

These dynamics seem to contribute to women's longer-term labor market outcomes. For example, a university field experiment found that one hour a week of single-sex college classes increased women's performance, especially for women at the lower end of the ability distribution (Booth, Cardona-Sosa, and Nolen 2018). Women in STEM doctoral programs with more women were also more likely to complete their degree on time (Bostwick and Weinberg 2018). Further, women in academia who coauthor with men get less credit for their work—and are consequently promoted at lower rates than similar men (Sarsons 2017).

In this paper, we contribute to several strands of this growing literature on gender and workplace dynamics. We conducted a multi-year field experiment with a top-10 U.S. undergraduate accounting program that was designed to identify the effects of work team gender composition on men's and women's performance, influence, and long-term labor market outcomes. Using survey, laboratory, and administrative data, we find that women who were randomly assigned to be the lone women on a team were rated as less competent and less influential than women in majority-women work groups. In a team laboratory task, they were less able to influence the outcome of the task. And while their grades were not affected by the gender composition of their group, we find that their long-term professional outcomes are affected. In particular,…

2. Experimental Design

We partnered with a top-10 undergraduate accounting program in the US and randomized the gender composition of students' assigned work teams during the fall semester of their junior year. Students are competitively admitted to the program at the end of their sophomore year. The fall semester is their first in the rigorous program. Assigned work teams are a critical part of their experience that semester: they attend class, work on cases and other course assignments, and study for exams with their five-person work teams. Program administrators have designed these work teams to replicate the work environment of many accounting firms, and faculty use these groups to train students on professional interactions. As is typical of work teams in the business world, students are instructed to make decisions together and respect each other's contributions. During the course of the semester, students complete both individual and group assignments, and they interact frequently with each other. The vast majority of groups meet together at least 2-3 days a week for an average of 10 hours a week. (see Table A1).[1]

The program enrolled 535 students (384 men and 151 women) in Fall semesters of 2016 and 2017.[2] Female students were randomly assigned to one of two groups: 1 woman and 4 men (1F) or 3 women and 2 men (3F). Because there were significantly more men than women in the program, men were randomized into one of three conditions: all five men (0F), 1F or 3F. Once individuals had been randomized into a condition, we randomly assigned them into groups. The program stipulated that the average GPA of each group should be similar and that there should only be one international student per group. Hence, we block randomized on these variables.

[1] This program is not unique in using work teams as a pedagogical and professionalization tool; many accounting programs and most MBA programs also assign students to similarly structured work teams.
[2] While we continued to collect data in the 2018-2019 academic year, the program directors decided to discontinue 1F groups resulting in a lower than expected sample size of 1F women.

We collected survey, laboratory, and administrative data to measure the outcomes of this randomization. Prior to group assignment, students took an extensive baseline survey. After group assignment, students took monthly surveys during fall semester that focused on their peer evaluations of their group members and their perceptions of group interactions.[3] At the completion of their second (senior) year of the program, we conducted an incentivized exit survey of all students to collect data on their internship experiences, post-graduation labor market outcomes, and graduate school plans. We also obtained administrative data on student grades and overall GPAs from the program.

In addition to the survey and administrative data, groups participated in an incentivized laboratory task at both the beginning and end of the semester. At the beginning of the semester, the task was the "Survival on the Moon" task[4], in which participants are given a list of 15 items and asked to rank them from most to least useful for survival on the moon, with expert answers being provided by NASA. At the completion of the semester, the task was a similar "Lost at Sea" task.[5] Both times, subjects completed the task in two stages. In stage 1, students first ranked the items on their own, giving us a measure of individual performance. In stage 2, participants then deliberated with their group members to create a group ranking of the items in the task. Immediately after groups submitted their rankings, each individual privately completed an exit survey about their group deliberation experience and voted for a spokesperson to present their group ranking to a panel of judges at a later date[6]. Each stage was presented to the subjects

[3] After groups were re-formed in Winter semester, we continued the monthly surveys.
[4] https://www.psychologicalscience.org/observer/nasa-exercise
[5] http://plcmets.pbworks.com/f/lost_at_sea.pdf
[6] Students were told that five groups would be selected at random to have an opportunity to have their spokesperson present their group ranking to the judges in 1-2 weeks for a chance to earn additional $50 for the best presentation. The groups were thus incentivized to select the most capable group member as their spokesperson.

separately and they did not receive any feedback on the task. We recorded these group deliberations[7] and incentivized lab performance in several ways: in addition to the $5 show-up fee, students could earn $50 for having the individual answer closest to the experts, $50 per group member for having the group answer closest to the experts, and another $50 per group for the best presentation by their chosen spokesperson.

Students were required to participate in the monthly surveys and laboratory exercises for course credit and for the purpose of an internal program evaluation. However, to be included in the analysis in this paper, they had to consent to allow their data to be used for research purposes.[8] Participants were informed that both the internal program evaluation and the research study would examine work team dynamics, and the experiment did not include any deception.

3. Sample and Balance Tests

In this section, we describe our sample and provide additional detail about the program. Admission to this program is highly competitive. Matriculated students typically have high GPAs (average 3.72) and considerable leadership experience. As reported in Table 1, the men and women in our sample both came to the program with impressive academic credentials, and there are no major differences between the men and women prior to their experience in the study. If anything, women appear to have more leadership experience.

[7] Recording equipment was 6-channel audio recording, which yields a separate, high-quality recording for each member of the group as well as records the group together.
[8] A small number of students did not consent to the use of their data, and their responses are not included in the results below.

Table 1: Mean Academic Performance and Leadership Experience

	Men	Women	Difference	P-value
High School Leadership	55.76	71.33	15.57	.001
Community Leadership	39.58	35.76	3.82	.415
Overall GPA	3.71	3.74	.02	.202
Prerequisite GPA	3.80	3.78	.02	.204

Table 2 reports our randomization checks by condition and gender. Across a wide variety of background characteristics, we see little difference between the men and women assigned to the various treatment conditions. A somewhat higher percentage of women in the 1F condition were international students, but in raw numbers, there were actually fewer international students in that condition than in the 3F condition. Together, the results in Table 2 affirm that randomization was successful, with conditions balancing across a wide variety of measures captured on the baseline survey and via administrative data.

Table 2: Randomization Balance Table

	1F Women	3F Women	*P-value	0F Male	1F Male	3F Male	*P-value
Average Age	22.49	22.72	.585	23.85	23.88	23.81	.761
Proportion White	.744	.870	.067	.939	.948	.917	.364
Proportion International	.205	.102	.101	.070	.058	.097	.292
(Number of International)	8	11	-	6	9	7	-
Proportion Married	.333	.231	.216	.341	.325	.361	.591
Average Prerequisite GPA	3.77	3.77	.946	3.80	3.79	3.81	.314
Average Key Prerequisite GPA	3.77	3.77	.988	3.77	3.77	3.79	.442
Parental Income Categories	3.79	3.82	.944	4.30	4.33	4.13	.442
Egalitarianism Index	.598	.570	.305	.558	.557	.564	.766
Individualism Index	.592	.607	.604	.681	.687	.662	.289
Pro-sociality Index	.766	.762	.829	.760	.762	.761	.972
Ambivalent Sexism Index	.595	.566	.260	.697	.708	.677	.069
Conflict Avoidance Index	.643	.630	.576	.555	.560	.546	.525
Opinionation Index	2.54	2.63	.522	2.83	2.84	2.76	.485
Responsibility Index	3.92	3.77	.348	4.16	4.18	4.11	.597
HS Leadership Experience	.692	.720	.749	.550	.558	.549	.899
Party Identification Scale	5.21	5.50	.287	5.85	5.84	5.85	.996
N	39	108		154	154	72	

Note: *P-values are between 1F and 3F Groups.

4. Results

In this section, we report the different outcome measures we collected through survey, laboratory, and administrative data. First, we focus on the different measures of influence and leadership. Across a wide variety of measures, we repeatedly find that lone women are significantly less influential than the women in the female-majority groups.

Monthly Peer Evaluations

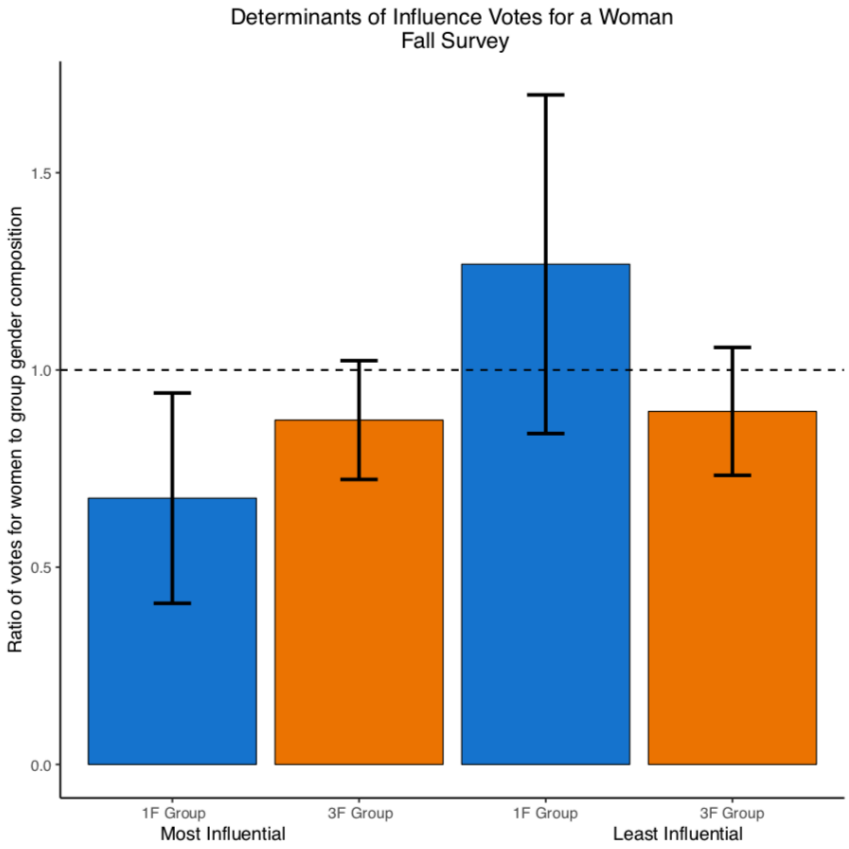
In each monthly survey, we asked respondents to identify the most and least influential member of their group during group discussion and collaboration. To analyze the responses, we computed the proportion of influence votes each participant cast for a woman over the course of the semester, then divided that number by the proportion of women in the group. This dependent variable is thus normalized across the conditions.[9] With this variable, a ratio of 1 means that the participant's proportion of votes for women was identical to the proportion of women in the group. Values less than 1 mean that the participant was disproportionately less likely to vote for a woman, and values above 1 mean that the participant was disproportionately more likely to vote for a woman. This measure thus approximates one potential standard of gender parity. It does not mean that influence is equally divided between men and women in the aggregate, but rather than influence mirrors the group's gender composition.

Figure 1 illustrates influence votes for women across treatments, aggregated across the course of the semester. As the figure shows, people in 1F groups were disproportionately less likely to vote

[9] This is an important adjustment, given that there are more women in 3F groups than 1F groups, so participants in those groups would be expected to cast more votes for women.

for a woman as the most influential member of their group. Similarly, we find evidence that women in 1F groups were disproportionately more likely to be considered as the least influential group member, though the confidence interval around that estimate is large. By contrast, people in 3F groups identified women as most and least influential at roughly proportionally to their representation in the group, though the point estimates are a little below the as-if-at-random rate. As reported in Table 3, the difference between experimental conditions is statistically significant at the 90 percent confidence level. This influence deficit is consistent with the previous work on the authority of token women in deliberating groups in the lab (Karpowitz and Mendelberg 2014).

Figure 1: Group Influence Votes – Aggregated Monthly Surveys



Note: 95% confidence intervals. Bars represent the average proportion of votes for women divided by the proportion of women in the group. (See Table 3 for full results)

Table 3: Determinants of Influence Votes for a Woman (Fall Survey)

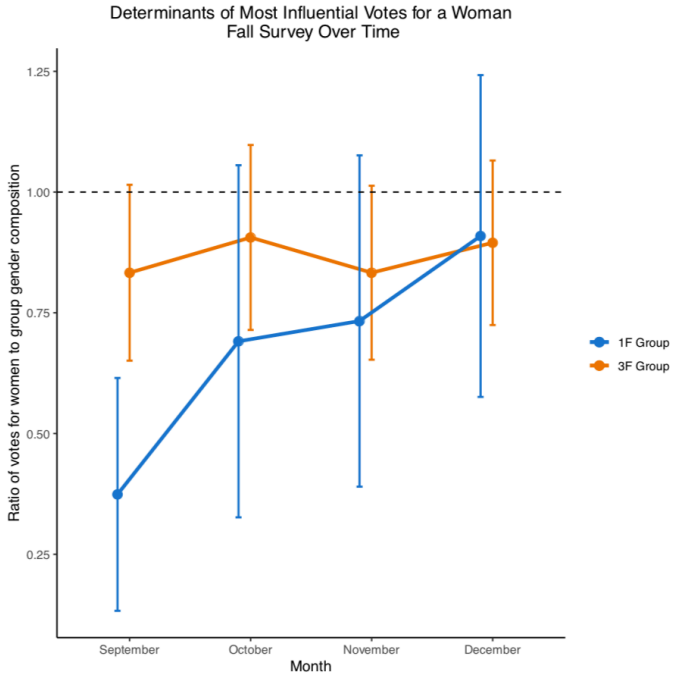
VARIABLES	(1) Most Influential	(2) Most Influential	(3) Most Influential	(4) Least Influential	(5) Least Influential	(6) Least Influential
1F Condition	-0.260*	-0.261*	-0.235	0.359	0.349	0.329
	(0.155)	(0.153)	(0.157)	(0.234)	(0.233)	(0.232)
Female	-0.157**	-0.132	-0.190*	-0.034	-0.048	-0.021
	(0.078)	(0.088)	(0.099)	(0.092)	(0.100)	(0.103)
International		-0.170	-0.207		0.145	0.194
		(0.271)	(0.277)		(0.364)	(0.386)
Age		0.021	0.021		0.011	0.011
		(0.025)	(0.024)		(0.028)	(0.027)
White		-0.119	-0.149		-0.206	-0.171
		(0.293)	(0.291)		(0.322)	(0.329)
Married		0.026	0.041		0.102	0.093
		(0.102)	(0.104)		(0.144)	(0.149)
Ambivalent Sexism			-0.603			0.328
			(0.393)			(0.533)
Constant	0.967***	0.582	1.002	0.916***	0.808	0.574
	(0.085)	(0.688)	(0.806)	(0.100)	(0.728)	(0.720)
Observations	367	367	366	363	363	362
R-squared	0.017	0.021	0.027	0.021	0.028	0.029

Note: Standard errors clustered by group. Dependent variable coded by taking the proportion of votes for women divided by the proportion of women in the group.
** p<0.01, ** p<0.05, * p<0.1

We find that this lower likelihood of granting women influence in 1F groups is not simply a function of women being less likely to vote for themselves. In fact, both men and women in these groups rate women as less influential, (Supporting Information Figures A1 and A2 and Tables A2 and A3). When we disaggregate by monthly survey, we see some evidence that women's influence deficit in 1F groups shrinks over time, though the confidence intervals are large (Figure 2). By the end of the semester, when considering the group generally, the influence deficit women faced in 1F groups appears to have evaporated, and there is little difference between the 1F and 3F conditions, though there is considerable variation across groups. Thus, when considering the group's interactions generally, we find that token women face a severe influence deficit in early interactions, though it is possible that repeated interaction improves the

situation, at least in some groups and with respect to a general assessment of influence in the group.

Figure 2: Group Influence Votes Over Time – Monthly Surveys



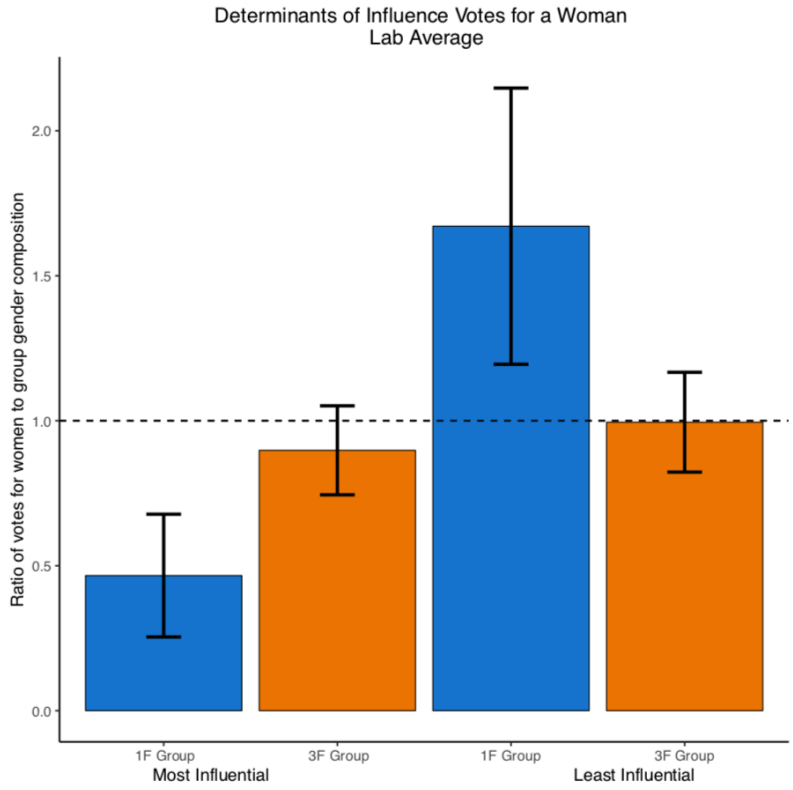
Note: 95% confidence intervals. Dependent variable coded by taking the proportion of votes for women divided by the proportion of women in the group.

Laboratory Task Influence Votes

In addition to the monthly peer evaluations, we asked a similar question to groups at the completion of their laboratory tasks, which were held at the beginning and at the end of the semester. Here, group members are instructed to evaluate their interactions on the task they had just completed instead of evaluating their group generally. We illustrate these results in Figure 3 below. The pattern is the same, but stronger. As seen in Figure 3, members of 1F teams were unlikely to see their female teammate as most influential in their laboratory task deliberations—they chose her about half as often as we would expect given the proportion of women in the group. They were also extremely likely to see her as the least influential person on the team. By

comparison, members of 3F teams chose women as most and least influential about as often as you would expect given their proportion in the group.

Figure 3: Group Influence Votes – Aggregated Lab Tasks



Note: 95% confidence intervals. Dependent variable coded by taking the proportion of votes for women divided by the proportion of women in the group.

As shown in Table 4, the differences between the 1F and 3F conditions are highly statistically significant and remain so even when we control for demographic characteristics and attitudes about women as measured by an ambivalent sexism index. In addition, when we disaggregate by the two lab tasks, we find no evidence that women's lot improves over the course of the semester. In contrast to the survey indicators, women in 1F groups experience a deficit of influence and authority at both the beginning and the end of the semester.

Table 4: Determinants of Influence Votes for a Woman (Lab Averaged)

VARIABLES	(1) Most Influential	(2) Most Influential	(3) Most Influential	(4) Least Influential	(5) Least Influential	(6) Least Influential
1F Condition	-0.477***	-0.453***	-0.451***	0.684***	0.666**	0.685**
	(0.136)	(0.138)	(0.137)	(0.257)	(0.256)	(0.259)
Female	-0.115*	-0.107	-0.097	0.020	0.066	0.005
	(0.067)	(0.071)	(0.080)	(0.130)	(0.151)	(0.157)
International		-0.009	-0.029		-0.274	-0.289
		(0.174)	(0.181)		(0.361)	(0.375)
Age		0.015	0.014		0.001	0.002
		(0.019)	(0.020)		(0.038)	(0.037)
White		0.249	0.235		-0.101	-0.120
		(0.153)	(0.160)		(0.362)	(0.375)
Married		-0.082	-0.080		0.055	0.065
		(0.098)	(0.100)		(0.177)	(0.180)
Difference from experts		0.011**	0.011**		-0.014*	-0.015*
		(0.005)	(0.006)		(0.008)	(0.008)
Talk Time		-0.007	-0.040		-1.610	-1.618
		(0.464)	(0.482)		(1.087)	(1.103)
Ambivalent Sexism			0.113			-0.671
			(0.411)			(0.635)
Constant	0.967***	-0.223	-0.274	0.983***	2.142*	2.678**
	(0.076)	(0.491)	(0.600)	(0.113)	(1.180)	(1.255)
Observations	373	373	371	373	373	371
R-squared	0.063	0.080	0.077	0.048	0.057	0.059

Note: Standard errors clustered by group. Dependent variable coded by taking the proportion of votes for women divided by the proportion of women in the group.
** p<0.01, ** p<0.05, * p<0.1

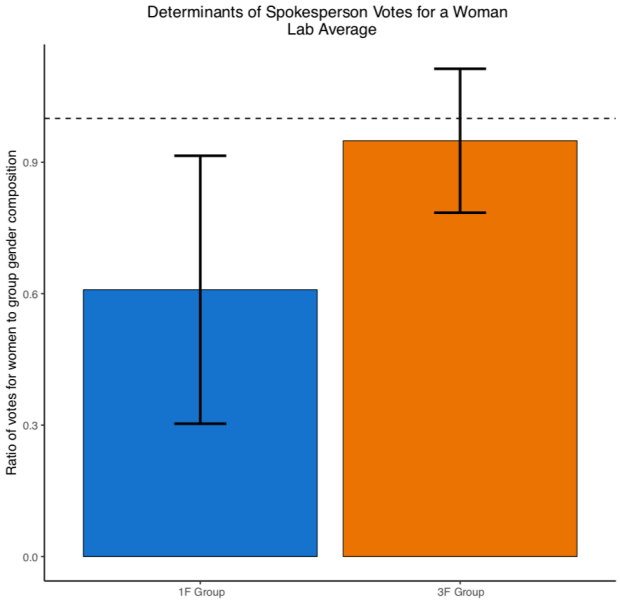
Laboratory Task Spokesperson Votes

The influence votes that we report from the monthly peer evaluations and the laboratory exit survey both indicate that token women are seen as less influential. Do these attitudes manifest themselves in behavioral outcomes? The design of our laboratory task offers several additional insights into this question. First, group members did not just vote on the most influential member of the group; they also made an incentivized choice - who to select as spokesperson for their group. The spokesperson would have a chance to earn another $50 for the team by doing a short presentation defending their group's decision. If the influence votes were perceived as a

low-stakes evaluation of group members, the selection of spokesperson was not. With an additional $50 at stake, each group member was incentivized to vote for the most competent member of the groups. In the spokesperson vote, participants voted privately and without notice or deliberation. Every group member was listed on the "ballot," so there was no nomination stage in which people could express that they were not interested in being a spokesperson.

As with the other measures of influence, we see treatment effects on the probability of voting for a woman as spokesperson. Figure 4 and Table 5 report these results. Once again, participants in 1F groups were disproportionately less likely to choose women relative to their proportion of group membership, while participants in 3F groups were not. As before, these results are robust to the inclusion of a wide variety of controls. The fact that women were also less likely to be chosen when the choice was consequential and incentivized suggests our attitudinal findings reflect real authority deficits.

Figure 4: Group Spokesperson Votes – Lab



Note: 95% confidence intervals. Dependent variable coded by taking the proportion of votes for women divided by the proportion of women in the group.

Table 5: Determinants of Spokesperson Votes for a Woman (Lab Averaged)

VARIABLES	(1) Spokesperson	(2) Spokesperson	(3) Spokesperson
1F Condition	-0.387**	-0.354**	-0.355**
	(0.170)	(0.170)	(0.172)
Female	-0.118	-0.093	-0.074
	(0.102)	(0.106)	(0.119)
International		-0.103	-0.117
		(0.176)	(0.183)
Age		0.034	0.034
		(0.022)	(0.022)
White		0.246	0.238
		(0.173)	(0.179)
Married		0.010	0.010
		(0.153)	(0.156)
Difference from experts		0.017**	0.018**
		(0.008)	(0.008)
Talk Time		-0.454	-0.481
		(0.731)	(0.740)
Ambivalent Sexism			0.208
			(0.319)
Constant	1.020***	-0.899	-1.033
	(0.100)	(0.908)	(0.946)
Observations	373	373	371
R-squared	0.026	0.051	0.050

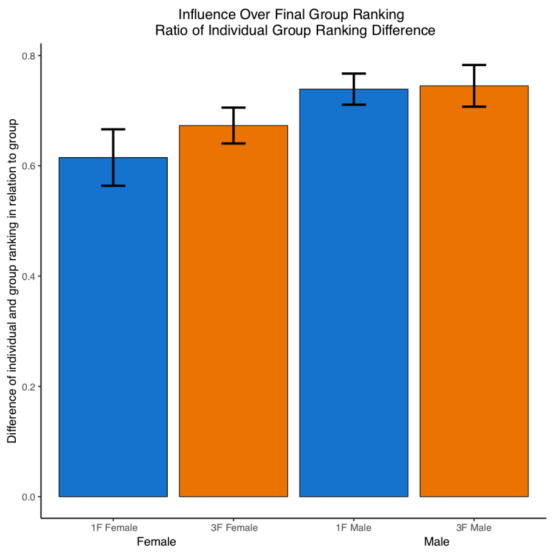
Note: Standard errors clustered by group. Dependent variable coded by taking the proportion of votes for women divided by the proportion of women in the group.
** p<0.01, ** p<0.05, * p<0.1

Empirical Influence on Laboratory Task Team Outcome

Beyond the votes from group members, we also have an additional, behavioral measure of influence. By comparing how individuals ranked items on their own to how the team collectively ranked items after they deliberated, we can glean an empirical measure of one's influence. We operationalize each individual's empirical influence as the absolute difference between an individual's ranking and the group ranking, and then normalize these scores for each group. In other words, the person in each group with the smallest absolute difference between their score and the group score would be given a ratio of 1 (most influential), with others scoring lower on the metric.

Overall, as shown in Figure 5 and Table 6, we find that women are significantly less likely to pull the team outcome toward their individual outcome than men. This is especially the case in 1F groups, as can be seen in Figure 5. Men in both conditions were more likely than women to exert actual influence, pulling the group toward their pre-deliberation preferences, and this difference is strongly statistically significant, as can be seen in Table 6. Women's influence deficit is strongest in groups with token women. In Model 2 of Table 6, we control for a variety of factors that could affect influence, including individual performance on the task. The inclusion of these factors absorbs some of the effect of the 1F treatment condition, though the overall gender disparity remains.

Figure 5: Group Ranking Ratio – Lab



Note: 95% confidence intervals. Individual group difference is made by taking the absolute value of the difference of each item ranking by the individual and the group. We then divide the minimum difference between the individual and group rankings by the individual ranking. (So, an individual group difference ratio of 1 would be the person with the lowest absolute difference.)

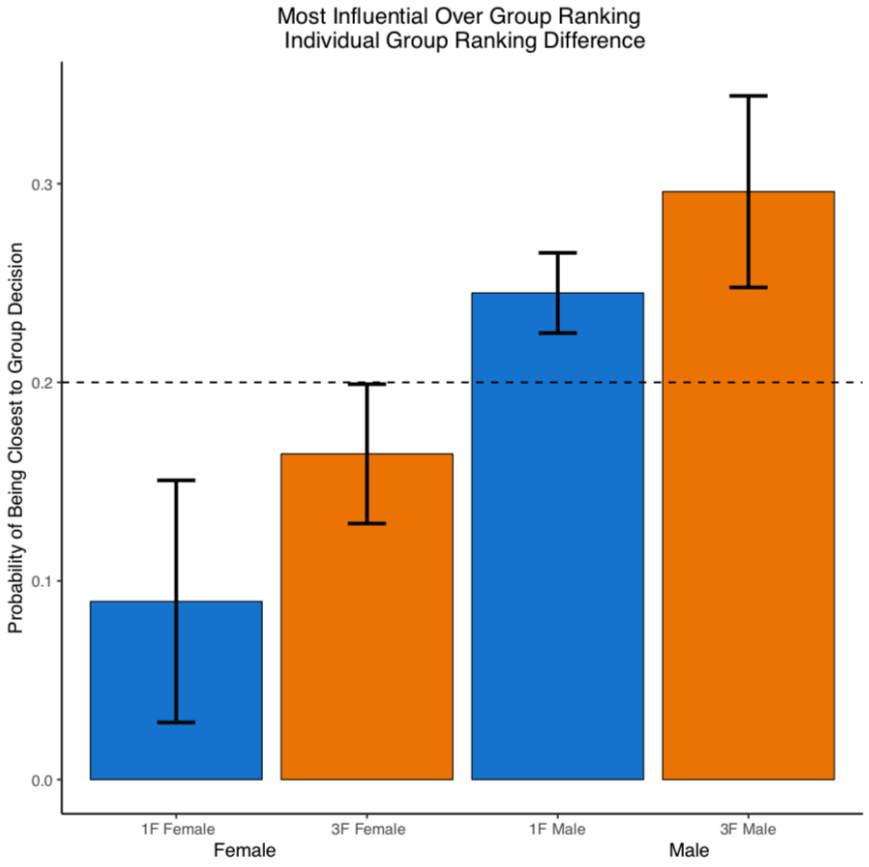
Table 6: Influence Over Final Group Ranking - Ratio of Individual Group Ranking Difference

VARIABLES	(1)	(2)
Female	-0.072***	-0.059***
	(0.020)	(0.019)
1F Condition	-0.007	-0.021
	(0.024)	(0.025)
1F X Female	-0.052*	-0.030
	(0.030)	(0.029)
International		-0.012
		(0.039)
Age		0.002
		(0.004)
White		0.024
		(0.037)
Married		0.039***
		(0.013)
Difference from experts		-0.004***
		(0.001)
Talk Time		0.560***
		(0.121)
Constant	0.745***	0.796***
	(0.019)	(0.134)
Observations	373	373
R-squared	0.061	0.188

Note: Standard errors clustered by group. Individual group difference is made by taking the absolute value of the difference of each item ranking by the individual and the group. We then divide the minimum difference between the individual and group rankings by the individual ranking. (So, an individual group difference ratio of 1 would be the person with the lowest absolute difference.)
** p<0.01, ** p<0.05, * p<0.1

We can sharpen these differences by exploring the likelihood of each individual member of the group having the greatest level of actual influence. In many ways, this analysis is analogous to the "most influential" votes we showed earlier. Here, however, the measure is whether the individual member of the group was most successful in persuading the other members of the group to follow his or her pre-deliberation preferences. These results are reported in Figure 6 and Table 7 below.

Figure 6: Best Group Ranking – Lab



Note: 95% confidence intervals. Individual group difference is made by taking the absolute value of the difference of each item ranking by the individual and the group. Those who had the lowest absolute difference were determined to be most influential on group decision.

Table 7: Most Influential Over Group Ranking

VARIABLES	(1)	(2)
Female	-0.132***	-0.115***
	(0.041)	(0.043)
1F Group	-0.051*	-0.082***
	(0.027)	(0.030)
1F X Female	-0.023	0.025
	(0.056)	(0.057)
International		-0.054
		(0.046)
Age		0.005
		(0.010)
White		0.002
		(0.050)
Married		0.017
		(0.032)
Difference from experts		-0.009***
		(0.002)
Talk Time		0.920***
		(0.264)
Constant	0.296***	0.481*
	(0.025)	(0.286)
Observations	370	370
R-squared	0.043	0.168

Note: Standard errors clustered by group. Individual group difference is made by taking the absolute value of the difference of each item ranking by the individual and the group. Those who had the lowest absolute difference were determined to be most influential on group decision.
** p<0.01, ** p<0.05, * p<0.1

The results are strikingly similar to what we found for influence votes in both peer evaluations and laboratory survey. Women in both the 1F and 3F conditions are less likely than men to exercise the most actual influence, though this pattern is especially pronounced among token women. The difference-in-differences does not achieve conventional levels of statistical significance, but the overall pattern is highly suggestive. When we add covariates, we see having pre-deliberation preferences that were closer to those of the experts and talking more frequently in the group discussion were both related to increased influence.

We explore the predictors of influence further by analyzing covariates separately by gender and experimental condition. These results are reported in Table 8. The dependent variable is whether or not the participant had the highest level of actual influence, though patterns are identical if we use the actual influence measure we reported in Table 6.

Table 8: Influence Over Final Group Ranking - Most Influential to Group Ranking

VARIABLES	(1) 1F Female	(2) 3F Female	(3) 0F Male	(4) 1F Male	(5) 3F Male
International	-0.166*	-0.088	0.019	-0.071	0.101
	(0.086)	(0.077)	(0.136)	(0.097)	(0.118)
Age	0.022	0.014	-0.013	0.006	-0.035**
	(0.028)	(0.009)	(0.017)	(0.012)	(0.017)
White	-0.070	-0.005	-0.137	0.147*	-0.073
	(0.077)	(0.083)	(0.120)	(0.082)	(0.128)
Married	-0.070	0.067	0.063	-0.004	0.088
	(0.074)	(0.060)	(0.066)	(0.053)	(0.081)
Difference From Experts	-0.488**	-0.243	-0.657***	-0.879***	-0.333
	(0.214)	(0.175)	(0.221)	(0.152)	(0.227)
Talk Time	0.064	0.844*	0.571	0.465	1.774***
	(0.469)	(0.435)	(0.368)	(0.480)	(0.436)
Ambivalent Sexism	0.345	-0.288**	0.259	-0.101	-0.211
	(0.220)	(0.142)	(0.217)	(0.182)	(0.346)
Constant	-0.233	-0.030	0.678	0.406	1.146**
	(0.576)	(0.258)	(0.448)	(0.354)	(0.447)
Observations	39	106	152	153	70
R-squared	0.226	0.174	0.140	0.184	0.223

Note: Standard errors clustered by group. Dependent variable is whether the individual had the smallest difference in individual to group ranking.
** p<0.01, ** p<0.05, * p<0.1

The patterns across the conditions are instructive. Token women were more likely to achieve actual influence when their pre-deliberation preferences were closer to those of experts (that is, when they were better at the task), but the coefficient on expertise for women in 1F groups is a only about half the size of the coefficient for men in the 1F condition. In other words, expertise appears to matter more for men than for women in that condition. In addition, talk time matters for women's influence in the 3F condition, but has no effect in the 1F condition. These differences are especially notable because token women do not have a large participatory deficit.

They speak about as much as we would expect, given their proportion in the group. Also recall that the women in the program are exceptionally strong students, so it is not surprising that they participate actively in group discussions. However, it appears that token women are not able to turn their active participation into influence. Women are more successful in doing so when they are in groups with other women.

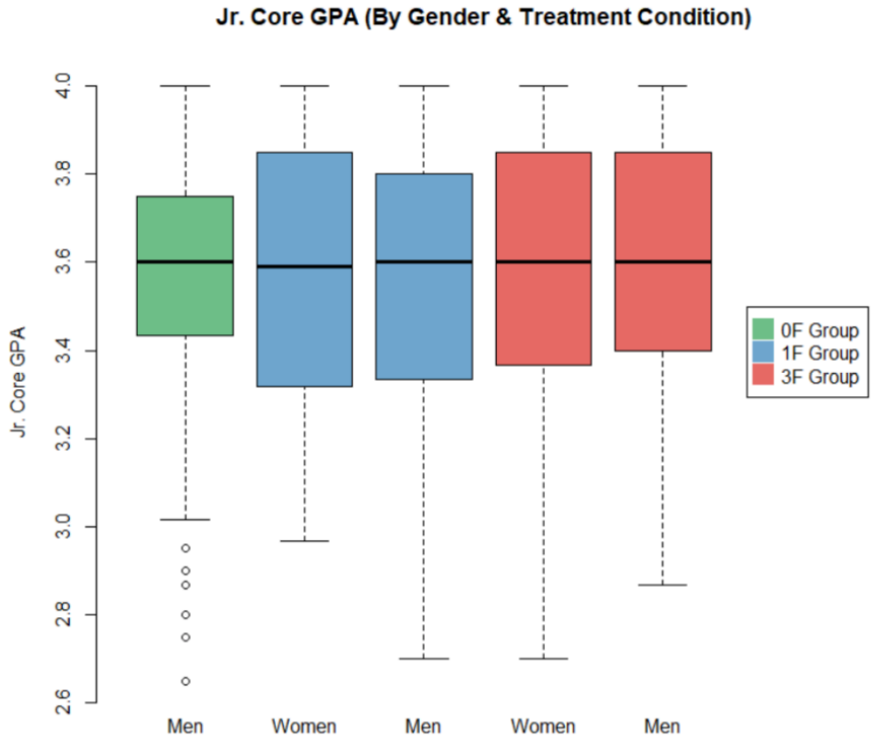
Taken together, an important pattern emerges. In 1F groups, influence appears to mostly be a function of competence at the task. From the perspective of efficient team production, that is to be expected. People who are better at a task should ideally be more influential. However, our findings suggest that men in these groups get almost twice as much return for their competence than women. By contrast, competence doesn't seem to affect empirical influence in 3F groups. In contrast, in 3F groups, influence seems to mostly be a function of talk time, but men in those groups get more than twice as much return for participation than women. It is clear that the gender composition of the team matters for how influence is allocated, and women face an uphill battle for influence in both kinds of groups vis-à-vis men.

Longer-Term Outcomes

We have shown that within the semesters we carefully studied these students, women in 1F teams face a persistent deficit in influence. In this section, we examine whether there is any evidence of longer-term effects of group gender composition. We collect both administrative and survey data on students' grades in the program, post-graduation plans and labor-market outcomes, including jobs and internship offers.

The first long-term outcome we examine comes from administrative data on students' final grades (GPA). As Figure 10 shows, there are no differences in students' grades in the program across gender or treatment conditions. This means that the treatment did not ultimately shift student's performance in the program, at least as measured by their grades. It is, of course, impossible to know whether this lack of difference is because grades capture something quite different than our measures of influence, or if perhaps women in 1F groups work harder to maintain the same grades as others.

Figure 10: Student Grades



In contrast to the findings about grades, we do find some evidence for treatment effects on post-graduation outcomes and career plans. In the exit survey that was administered in April and May of the students' senior year (more than a year after the experimental intervention), we asked both men and women to report their career plans by telling us about their internships, job applications, job offers, and graduate school aims[10]. Women in 1F and 3F groups had a similar number of internships, but women in the 1F groups applied for significantly fewer jobs: .30 versus 1.45 ($p = 0.033$). This may be because 1F women seem to choose a different career path than 3F women. While we lack statistical power to detect the effect with confidence, Table 10 presents suggestive evidence that women assigned to 1F teams are 14.8 percentage points more likely to plan to attend graduate school at the end of their senior year ($p = .129$). Interestingly, they are not less likely to have full-time work plans—a subset of the 1F women plan to attend graduate school and work full time after graduation.

Table 10: Career Plans (Exit Survey)

	1F Women	3F Women	*P-value	0F Male	1F Male	3F Male	*P-value
Number of Internships	1.06	0.98	.613	0.87	0.89	0.99	.411
Number of Jobs Applied For	0.29	1.45	.033	2.45	2.34	1.32	.136
Percent Interviews by Applications	100%+	81.5%	.116	100%+	77.5%	100% +	.026
Number of Job Interviews	0.65	0.96	.381	1.53	1.32	1.41	.847
Percent Offers by Interviews	78.9%	55.4%	.110	62.3%	75.3%	73.4%	.846
Number of Job Offers	0.44	0.54	.591	0.71	0.77	0.70	.673
**Grad School Plans	73.5%	58.7%	.129	59.2%	57.9%	54.0%	.598
**Full-Time Work Plans	32.4%	37.0%	.635	45.4%	42.8%	41.3%	.843
N	34	92		130	145	63	

Note: 95% confidence intervals. *P-values are between 1F and 3F Groups. **Students were asked to select all immediate plans they had after graduation (Work full-time, work part-time, attend graduate school in summer or fall, take time to prepare for graduate school, be a homemaker or stay at home parent or other)

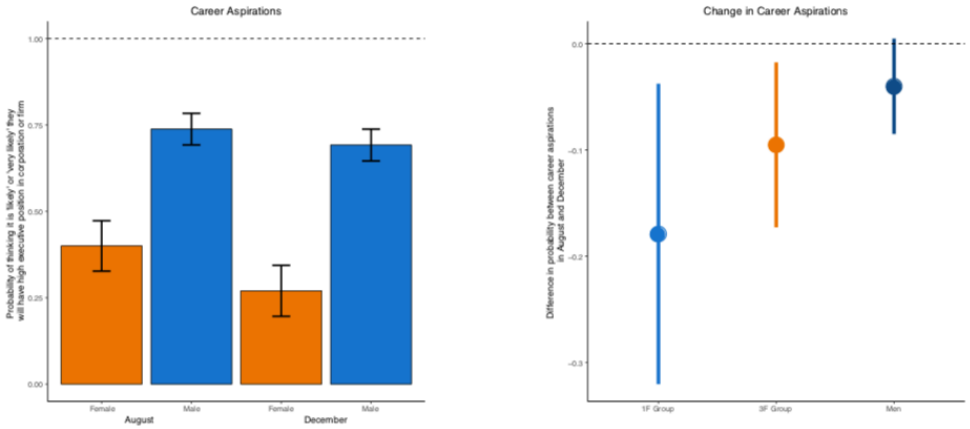
There are several possible explanations for the finding that random assignment to either a 1F or 3F group has a lasting effect on women's career trajectories. Attending graduate school is seen

[10] While we incentivized the completion of the survey with a $7 payment, not all the students responded to this optional survey, resulting in a smaller sample size. We compare the demographics of the students who responded with the rest of our subject pool and find no statistically significant differences.

as an ambitious and prestigious choice.[11] So, it is possible that formative experiences in a male-dominated environment significantly increased these women's career ambitions. Yet, we have some evidence that there may be a different mechanism at play. First, 1F women are attending graduate school at rates substantially higher than the men, so they do not appear to be simply acculturated into masculine career ambition. Second, in August of the students' junior year (before groups were assigned) and again in December (at the end of the first semester), we asked students how likely they would be to "have a position as partner or director of an accounting firm or other high executive position in a corporation or firm" in the next 15 years. Figure 11 illustrates the responses to this question. In both August and December, women were significantly less likely than men to envision themselves as corporate or accounting executives. Over the course of the semester, both men and women became somewhat *less* likely to see themselves in that role, though the decline was greater among men than among women. And although the differences across experimental conditions are not statistically significant, the decline is largest among women in the 1F condition, as can be seen in the right panel of Figure 11. These data suggest that 1F women's decisions to attend graduate school may not stem from greater ambition, but instead from lack of confidence about future career prospects.

[11] This information comes from conversations with program directors.

Figure 11: Career Aspirations Over Time



Note: 95% confidence intervals. Students were asked "In the next 15 years, how likely do you think it is that you will have a position as a partner or director of an accounting firm or other high executive position in a corporation or firm?" Those who responded with "Likely" or "Very Likely" are coded as a 1 and all other responses are coded as a 0. The difference is found by subtracting the August score from the December score.

5. Discussion and Conclusion

In this paper, we describe a field experiment designed to study whether gender composition of work teams influences the ways in which team members are evaluated. In partnership with a top-10 accounting program in the U.S., we randomized female students into 5-person teams consisting of either one woman or three women. Students interacted extensively in their teams over the course of a semester (4 months) before the groups were re-assigned. We collect administrative, survey, and lab data on these students' outcomes in the program, their experience in teams and post-graduation goals and outcomes. In addition to unique field and lab data on students' participation and influence in their groups, we also collect monthly data allowing us to consider changes over time.

Across multiple measures of influence, we find strong evidence that token women experience a deficit of authority and influence. Women in the 1F condition in our study were substantially less

likely to be judged as influential by the other members of the group, relative to their presence in the group, and these deficits were most pronounced in the lab, where we asked group members to evaluate contributions to a concrete decision they had made just moments before. While we saw some evidence that women's influence increased over time in the monthly survey measures, there is no change from the beginning to the end of the semester in the lab. Moreover, in the lab, women in the 1F condition were less likely to be chosen to represent the group as spokesperson, a choice that was made meaningful by a substantial incentive. We also find that token women were less likely to exert actual influence, as assessed by their ability to pull the group toward their pre-deliberation preferences. The differences between conditions was not statistically robust in every case, but the broad pattern across different measures was consistent. Token women were less likely to be granted authority by the other members of the group – often less than women in the 3F condition, but nearly always less often than we would expect if influence were distributed equally in the 1F groups. In monthly survey responses, we find that this pattern was similar in judgments about competence, but very different in judgments about women's supportiveness of group efforts. Thus, token women were less likely to be seen as influential leaders, but group members had little trouble granting women a more supportive role in the group.

Further, we examine whether there is any evidence of longer-term effects of group gender composition. We first analyze administrative and survey data on students' grades and post-graduation outcomes, including graduate school plans, jobs and internship offers. While we find no difference in grades of 1F and 3F women, we do find differences in their post-graduation career plans. In particular, women in 1F groups report a significantly higher likelihood of

attending graduate school, as opposed to pursuing a job immediately after graduation. Though we cannot be sure of the mechanism that leads lone women to pursue graduate school at a higher rate than any other group, we have suggestive evidence that it is more likely the result of decreases in confidence in career trajectory than increases in career ambitions.

All of this leads us to conclude that who is granted authority is highly context-dependent. Although women entered this accounting program with identical grades to men and more leadership experience, their ability to be seen as influential and competent leaders in their teams was deeply shaped by whether or not they were a lone woman. This finding has important implications for how we think about workplace team dynamics. Evaluations of women's performance—as well as their ability to shape team outcomes—is endogenous to the gender composition of the team.

The most important implication is that women in male-dominated workplace environments are at a distinct disadvantage. Since performance reviews frequently rely on subjective evaluations of contributions, competence, and leadership, women in these settings are likely to have a harder time gaining recognition and advancement. Women have reported these kinds of workplace difficulties for decades (Kanter 1977). However, many have been skeptical about their existence or importance without clear causal evidence.[12] We have now provided experimental evidence of these dynamics in long-term work teams. Across a wide variety of outcomes, we find that influence becomes highly gendered when men are in the majority. Workplaces that care about

[12] This skepticism about women's complaints of gender bias may, of course, be rooted in the same patterns of socialization that lead team members to discount the authority of women teammates.

gender parity should take note—as should any workplace that is interested in efficiently recognizing and using all workers' competencies.

WORKS CITED

AAUW. 2016. “Barriers and Bias: The Status of Women in Leadership.” *AAUW: Empowering Women Since 1881*. https://www.aauw.org/research/barriers-and-bias/ (March 31, 2019).

Altonji, Joseph G., and Rebecca M. Blank. 1999. “Chapter 48 Race and Gender in the Labor Market.” In *Handbook of Labor Economics*, Elsevier, 3143–3259. http://www.sciencedirect.com/science/article/pii/S1573446399300390 (March 31, 2019).

Babcock, Linda, Maria P. Recalde, Lise Vesterlund, and Laurie Weingart. 2017. “Gender Differences in Accepting and Receiving Requests for Tasks with Low Promotability.” *American Economic Review* 107(3): 714–47.

Booth, Alison L., Lina Cardona-Sosa, and Patrick Nolen. 2018. “Do Single-Sex Classes Affect Academic Achievement? An Experiment in a Coeducational University.” *Journal of Public Economics* 168: 109–26.

Born, Andreas, Eva Ranehill, and Anna Sandberg. 2018. *A Man’s World? The Impact of a Male Dominated Environment on Female Leadership*. Rochester, NY: Social Science Research Network. SSRN Scholarly Paper. https://papers.ssrn.com/abstract=3207198 (March 31, 2019).

Bostwick, Valerie K, and Bruce A Weinberg. 2018. *Nevertheless She Persisted? Gender Peer Effects in Doctoral STEM Programs*. National Bureau of Economic Research. Working Paper. http://www.nber.org/papers/w25028 (March 31, 2019).

Bureau of Labor Statistics. (2017). Employment status of the civilian noninstitutional population by age, sex, and race. *Labor Force Statistics from the Current Population Survey*. Retrieved from http://www.bls.gov/cps/cpsaat03.htm

Eckel, Catherine C., and Philip J. Grossman. 2005. “Managing Diversity by Creating Team Identity.” *Journal of Economic Behavior & Organization* 58(3): 371–92.

Isaksson, Siri. 2018. “It Takes Two: Gender Differences in Group Work.” https://www.dropbox.com/s/hafhh70p3li9tij/1120jmp.pdf?dl=0 (March 31, 2019).

Jehn, Karen A., Gregory B. Northcraft, and Margaret A. Neale. 1999. “Why Differences Make a Difference: A Field Study of Diversity, Conflict and Performance in Workgroups.” *Administrative Science Quarterly* 44(4): 741–63.

Kanter, Rosabeth Moss. 1977. “Some Effects of Proportions on Group Life: Skewed Sex Ratios and Responses to Token Women.” *American journal of Sociology* 82(5): 965–90.

Lazear, Edward P., and Kathryn L. Shaw. 2007. “Personnel Economics: The Economist’s View of Human Resources.” *Journal of Economic Perspectives* 21(4): 91–114.

Preece, Jessica Robinson. 2016. “Mind the Gender Gap: An Experiment on the Influence of Self-Efficacy on Political Interest.” *Politics & Gender* 12(1): 198–217.

Sarsons, Heather. 2017. “Recognition for Group Work: Gender Differences in Academia.” *American Economic Review* 107(5): 141–45.

Thomas-Hunt, Melissa C., and Katherine W. Phillips. 2004. “When What You Know Is Not Enough: Expertise and Gender Dynamics in Task Groups.” *Personality and Social Psychology Bulletin* 30(12): 1585–98.

Yang, Yang, Nitesh V. Chawla, and Brian Uzzi. 2019. “A Network’s Gender Composition and Communication Pattern Predict Women’s Leadership Success.” *Proceedings of the National Academy of Sciences* 116(6): 2033–38.