This version: March 21, 2011

Gender Bias in NIH Peer Review: Does it Exist and Does it Matter?

Danielle Li

Abstract

This report examines the extent to which gender influences peer review at the NIH. Examining the percentile rankings assigned to over 45,000 male and female R01 applicants, I find that female investigators receive on average a half percentile worse ranking than male investigators with similar qualifications and similar future grant performance. Gender bias reduces the number of female applicants who are funded by 0.49 to 3.44 percent. Analysis of study section composition reveals that the presence of female reviewers attenuates bias, suggesting that gender balance on study sections can improve peer review.

1 Introduction

There has been longstanding policy concern about the lack of women in the upper tiers of the biomedical sciences; although women currently make up half of doctorates in the life sciences, they account for only a quarter of tenured or tenure-track faculty in these areas. This cannot be explained by cohort effects alone: women are disproportionately represented among non-tenure track positions and experience more attrition than male scientists at each stage of the post-doctorate career ladder [1] [2] [3]. Many explanations have been put forth to explain this "leaky pipeline" including preferences for family-friendly careers, dual-career concerns, and lack of mentorship [1] [3] [4].

In this report, I consider the role of research grants. It is difficult to overstate the importance of NIH funding in the academic life sciences. With an annual budget of 30 billion dollars, the NIH is the largest supporter of academic biomedical research in the US and success as an independent investigator depends on the ability to obtain these funds; without it, most academic life scientists would be unable to pursue their own research agenda, maintain a lab, obtain tenure, or even pay their own salaries.

At the same time, success rates for NIH grants have fallen to under 18 percent for new applications, raising concerns that scarcity and uncertainty in funding may deter promising graduate

students and postdocs from pursuing academic research careers. These concerns may be exacerbated for female scientists who are already less likely to pursue the senior-level academic positions that rely most on external support. In this context, gender bias in the peer review process, either perceived or real, is potentially high-stakes and may contribute to the attrition of female scientists.

NIH peer review for most R01 applications works as follows: applications are assigned to a study section for review and to an Institute and Center (IC) for funding. Study sections assess the scientific merit of applications by assigning them a priority score, which, during the period my data come from, ranged from 1.0 for the best application to 5.0 for the worst, in increments of 0.1. This priority score is converted into a percentile from 1 to 99, where a percentile reflects the percentage of applications from the same study section and reviewed in the same year that received a better priority score. However, for ease of exposition and intuition, I report percentiles to mean the percentage of applications that are worse, so that higher percentiles are better. In most cases, proposals are funded in order of their percentile until the designated funding IC reaches its payline.

Although men far outnumber women among NIH grant recipients, success rates for male and female applications are relatively similar. Previous studies have taken this parity to mean that gender is unlikely to play a role in the grant review process [2] [5] [6]. This, however, is not a robust conclusion without more information on the qualifications of applicants. If female applicants are on average more qualified than male applicants, then parity in success rates implies discrimination against women. If, on the other hand, female applicants are less qualified, then parity implies discrimination in favor of women.

So far, no studies carefully account for the qualifications of grant applicants or measures of application quality in examining the role of gender in grant provision. In this report, I use NIH administrative data on individual grants and their percentile rankings to better understand how the NIH peer review process treats applications from female investigators. For 46,093 successful R01 grant applications from 1992 to 2007, I match these NIH data with PI degrees, institutional affiliation, NIH funding history, past publication history, and gender. I take the additional step of tracking the performance of each funded grant by matching it to the future publications it generates as well as to the citations accruing to those publications.

My sample is limited to funded grants and, as such, my analyses may not be representative of gender gaps in rankings among unfunded grants. The benefit of examining funded grants, however,

is that I am able to obtain measures of PI qualifications observable to study sections as well as measures of future grant performance that are unobserved at the time of review. These measures of project quality are not available for unsuccessful applications because future performance is not observed if the project is not funded.

Using these detailed measures of grant quality, I compare the percentiles assigned to grants whose PIs differ in gender, but which are evaluated in the same study section meeting, have similar publication and funding histories, and, moreover, eventually produce similarly cited research. I then quantify the effect of gender bias on the number of women who are funded. For approximately half of my observed grants, I also collect data on the gender of the NIH study section members to whom the grant was assigned for review. For this subsample, I ask whether the gender composition of the study section influences how female applicants are evaluated.

My results indicate that women face larger hurdles than men, especially in the R01 renewal process. For new R01s, women receive on average a third of a percentile worse ranking than male investigators with similar qualifications, but this gap rises to two thirds for competing applications. I find that gender bias reduces the number of funded women by 0.49 to 3.44 percent. Analysis of study section composition reveals that the presence of female reviewers attenuates gender bias; study sections appear to be unbiased when about a third of their members are women.

2 Identifying Gender Bias

Figures 1 and 2 show that despite substantial gains, women comprised only a third of R01 awardees as of 2007 and, conditional on being funded, received worse percentile rankings. At the same time, however, female R01 awardees also have weaker publication records, as shown in Figure 3. Together these patterns are inconclusive: women receive fewer grants and worse rankings, but this disparity may be reflective of an underlying lack of highly qualified female applicants as opposed to any systemic bias.

Identifying bias from aggregate statistics is difficult because aggregate statistics conflate differences in how applicants are treated, conditional on quality, with unobserved differences in applicant quality. Instead, I compare the percentiles assigned to grants awarded to male and female scientists to the same study section. In the raw comparison, women receive percentiles that are on average

0.461 ($P = 0.006$) and 0.854 ($P < 0.001$) worse than men for new and competing respectively (see Figure 4). Some of this gap, however, may be accounted for by other observable differences between applicants; female investigators are also younger, have fewer high-impact publications, and are less likely to be affiliated with elite institutions. Restricting my comparison to male and female applicants who also have similar past publications, grant histories, and affiliations reduces the gender gap in assigned percentiles to 0.324 ($P = 0.055$) for new grants and 0.634 ($P = 0.001$) for renewals (See Supporting Materials for a full list of controls).

There are two classes of possible explanations for this remaining disparity. The first is that women receive worse rankings for reasons not justified by the quality of their research. The second is that women receive worse rankings because they are less qualified along dimensions that I do not observe. If, for instance, women are less likely to propose innovative methods, then the gender penalty I find may reflect this and not gender per se. More generally, study sections do not observe everything about an application's quality and may instead attempt to infer quality based on what they do observe. In this case, because female scientists tend to have fewer qualifications along many observable dimensions, committee members may assume that even though two applicants have similar observable qualifications, the female applicant may still be less qualified on some unobserved dimension.

These cases can be distinguished from each other using measures of future grant performance. Imagine that each grant has some quality that study sections care about, but which they do not perfectly observe. Were study sections using gender to infer quality, then gender should matter for percentile rankings only insofar as it is correlated with quality; that is, gender should not matter if the study section thinks that two applicants have the same quality. Any remaining effect of gender is then attributable to bias or some inference about quality that is incorrect (See Supporting Information for details). I test this hypothesis by controlling for the future grant performance and find a gender gap of 0.322 percentiles ($P = 0.056$) that is due to bias for new grants and 0.597 ($P = .002$) percentile for competing grants. Female investigators on average receive the same percentiles as comparable men even when their research eventually performs better and this is particularly salient for competing grants. This is evidence that study sections underestimate the quality of female investigators. These gender gaps are graphed in Figure 4.

3 How Large Are These Effects?

What are the consequences of a half percentile bias against women for the number of women who receive funding? The answer depends on how much bias exists near the payline. A small bias may have large consequences for who gets funded if it pushes an applicant across the payline while even large biases may not matter if they occur far from the payline. Figure 2 indicates that women are relatively over-represented at worse percentile scores; in this case, small differences in scores may lead to changes in the number of women who are funded. I quantify the effect of bias using a counterfactual analysis to estimate how many women would have funded had gender not played a role in funding.

Historically, the number of R01 equivalent awards per year have ranged between 6,000 to 7,500. Thus, for the sample of funded grants I observe, I consider instead which of those grants would have been funded if the number of NIH grants were reduced by 5, 10, 15, and 25 percent. For each of these cases, I construct counterfactual percentile rankings of grants under a gender neutral review. These re-rankings allow me to re-order grants and create a different allocation of grant winners under each assumption. I compare the number of women who are funded under each of these different portfolios. These calculations indicate that gender bias leads to a 0.49 to 3.44 percent decline in the number of funded women, depending on what percent of funding is being cut (Figure 6). Details are provided in the Supporting Information.

4 How Can Study Section Performance Be Improved?

The results I find may not be representative of all study sections. I obtain attendance rosters for 2,290 study section meetings from 1992-2005 and match grant applications to the study section in which they were reviewed in order to examine whether the presence of female study section members impacts the assessment of female applicants.

In this subsample, there is no overall gender bias in review but, for competing applications, the presence of women attenuates gender bias. In study sections with no female reviewers, women face a 2.025 ($P = 0.015$) percentile bias in competing applications, but that each additional 10 percentage point increase in the proportion of female reviewers decreases this bias by 0.601 ($P = 0.035$) percentiles. Gender bias is neutralized when study sections are about one third female, which is

nearly the average in my sample of standing study sections. These results are consistent with recent research demonstrating that female representation is correlated with increased group performance [9].

5 Conclusion

My results show that women on average receive a half percentile worse rank than those of similarly qualified men, leading to a 0.49 to 3.44 percent reduction in the number of female investigators who are funded. This is evidence that study sections make systematic mistakes when judging the quality of female applicants relative to their male peers. The true cost of gender bias, however, includes both the actual loss of female investigators at the margin of funding as well as the potential that perceptions of bias may discourage female scientists from pursuing academic research careers. The latter effect may be especially salient since other research suggests that they are already more reluctant to pursue these careers due to concerns about work-life balance [3] [1]. I find that these problems are attenuated by the presence of more female reviewers on study sections. In particular, bias against female applicants is neutralized when a third of study section members are women. This is true for half of study sections in my sample and, moreover, 75 percent of study sections are at least 20 percent female and almost no sections are less than 10 percent female. NIH efforts to promote the representation of women on study sections appears to be an important step toward ensuring that grant review is both fair and perceived as fair.

References

[1] NSF "Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering." Arlington, VA, USA: National Science Foundation (2006).

[2] T. Ley, B. Hamilton, "The Gender Gap in NIH Grant Applications." *Science*. **322**, 1472 (2008).

[3] E. D. Martinez, J. Botos, K. M. Dohoney, T. M. Geiman, S. S. Kolla, A. Olivera, Y. Qiu, G. V. Rayasam, D. A. Stavreva, O. Cohen-Fix, "Falling off the academic bandwagon." *EMBO Reports*. **8**, 977 (2007).

[4] S. Ceci, W. Williams, "Understanding Current Causes of Women's Underrepresentation in Science." *PNAS*. **108**,8 (2011).

[5] RAND "Is there a Gender Gap in Federal Grant Programs?" Brief No: RB-9147 (2005).

[6] P. Leboy, "Fixing the Leaky Pipeline." *The Scientist*. **22**, 67-70 (2008)

[7] F.D. Blau, J. M. Currie, R. T. A. Croson, D. K. Ginther, "Can Mentoring Help Female Assistant Professors? Interim Results from a Randomized Trial." *American Economic Review*, **100**, 2 (2010).

[8] F. Fang, A Casadevall. "NIH Peer Review Reform—Change We Need, or Lipstick on a Pig?" *Infection and Immunity*. **77**,3 (2009).

[9] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, T. W. Malone. "Evidence for a Collective Intelligence Factor in the Performance of Human Groups" *Science*. **330** 6004 (2010).

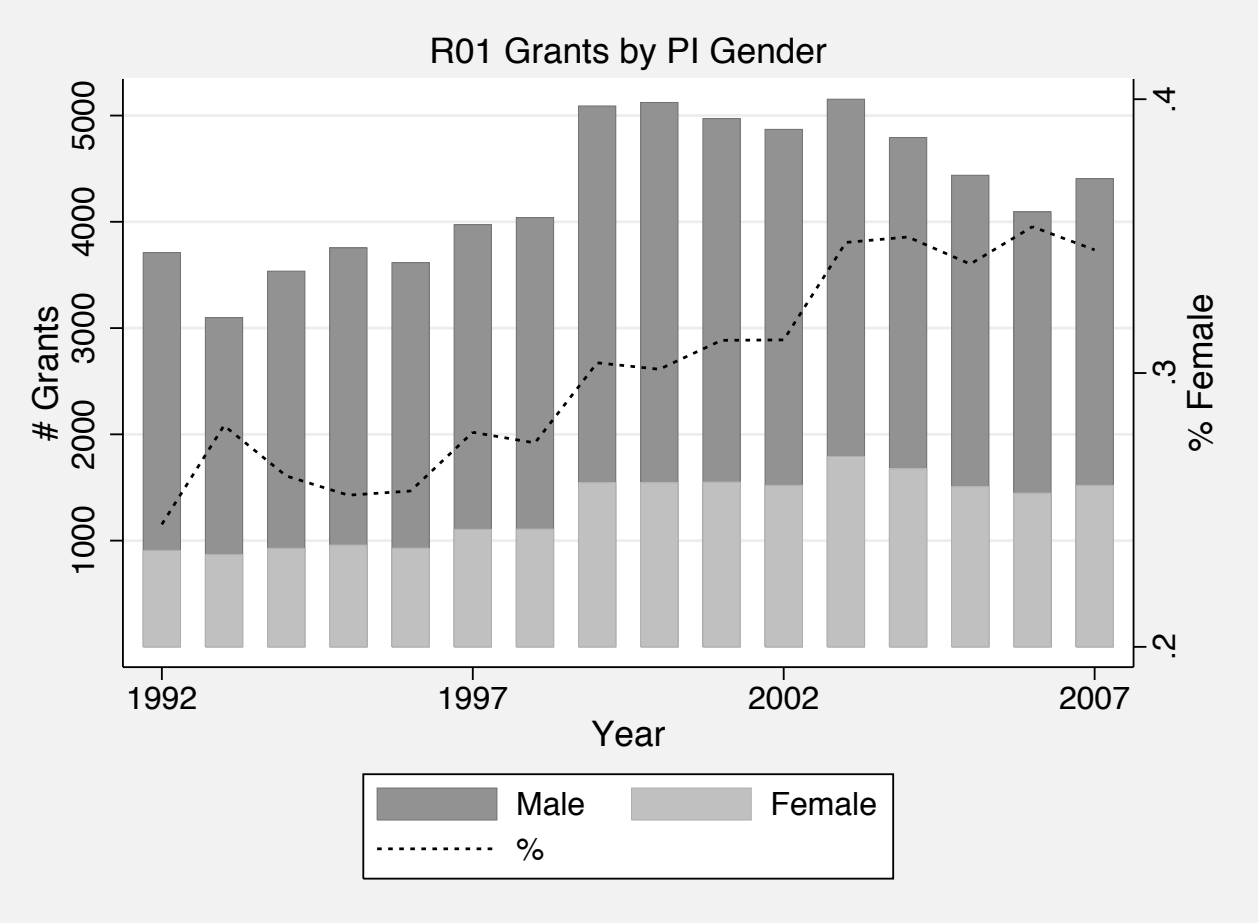

Figure 1: Representation of female investigators among R01 grantees has risen over time, but still remains low. Number of grants refers to the number of new or competing grants funded per year.

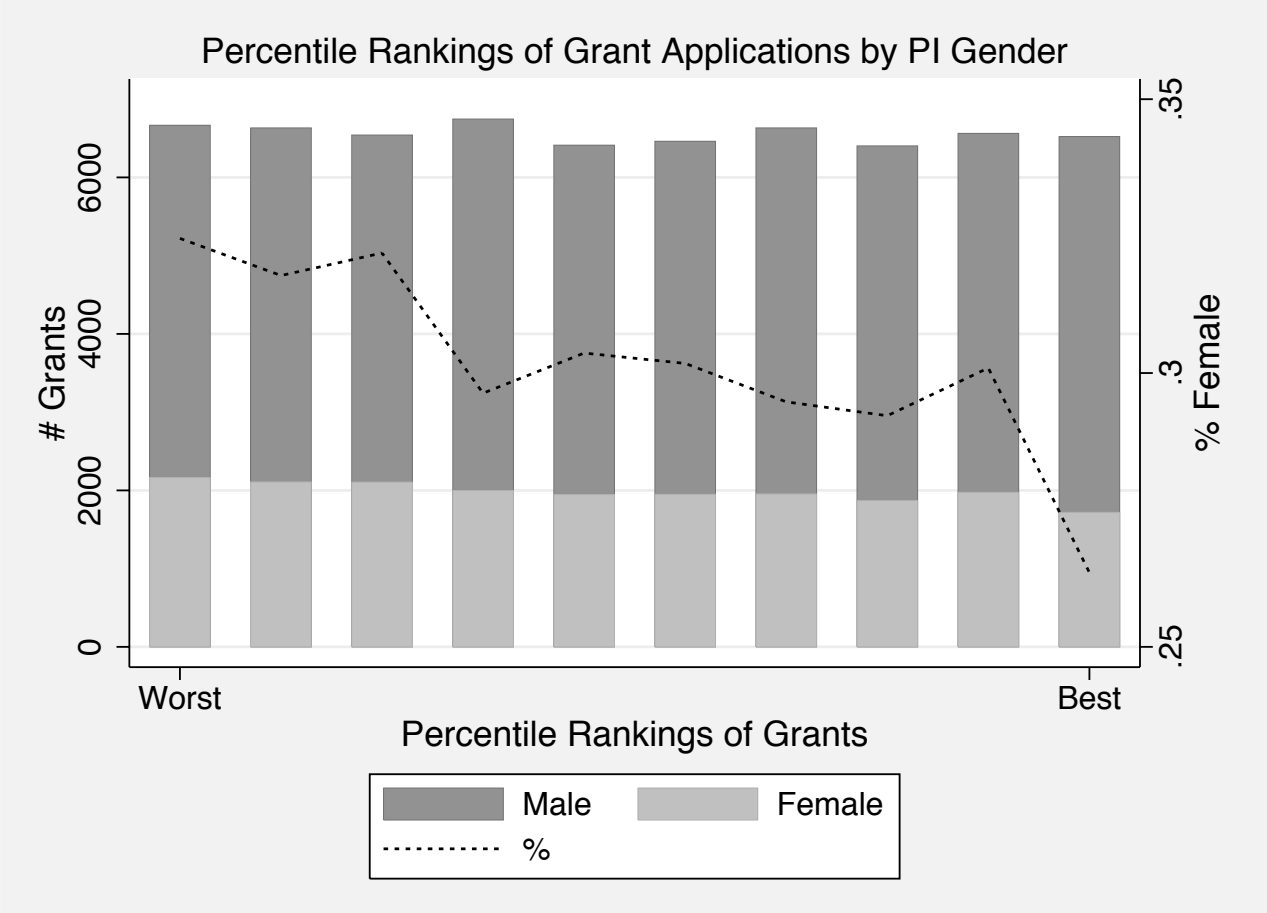

Figure 2: Female investigators are more represented among funded R01 grants with worse percentile rankings.

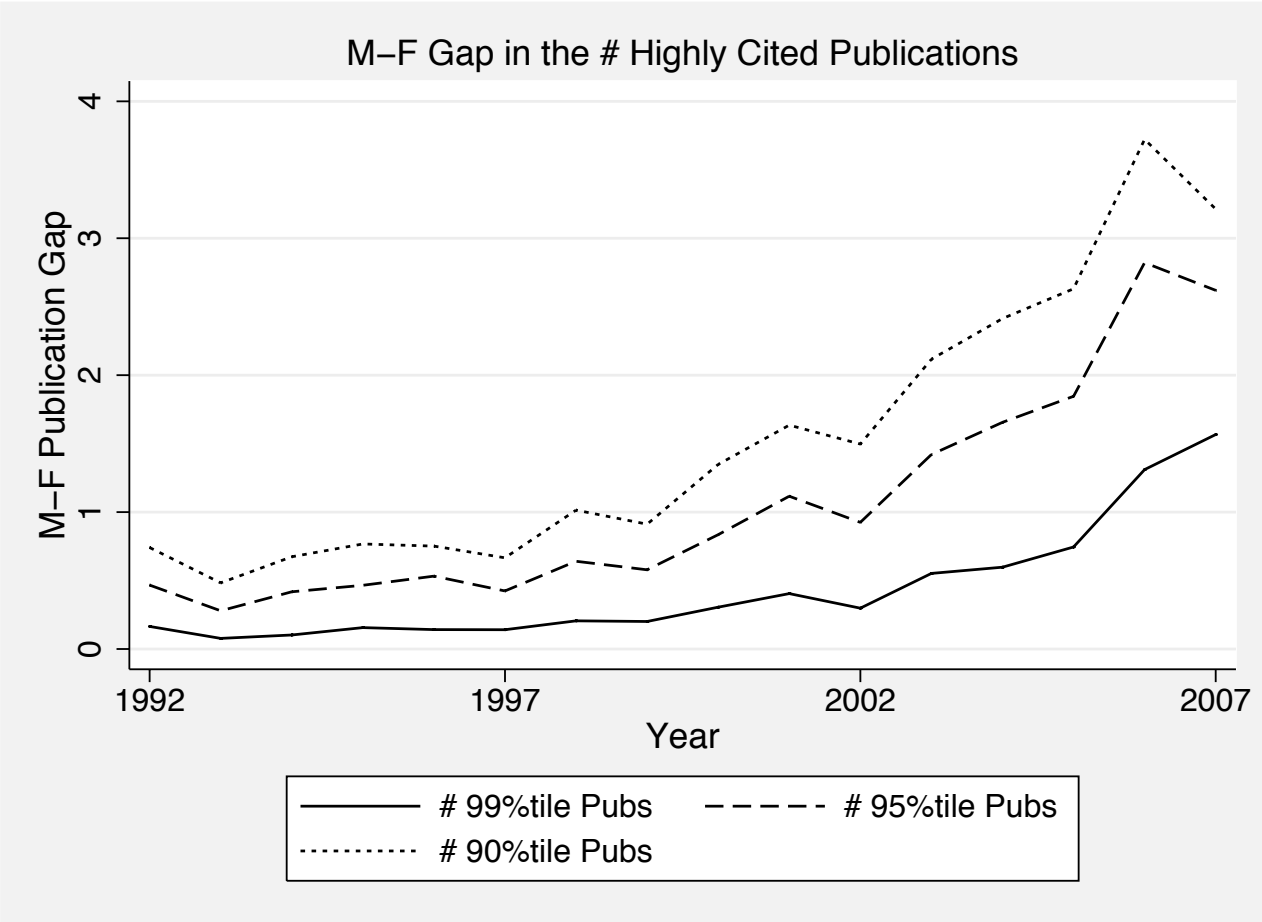

Figure 3: Female investigators have fewer highly cited publications in the five years prior to their grant application.

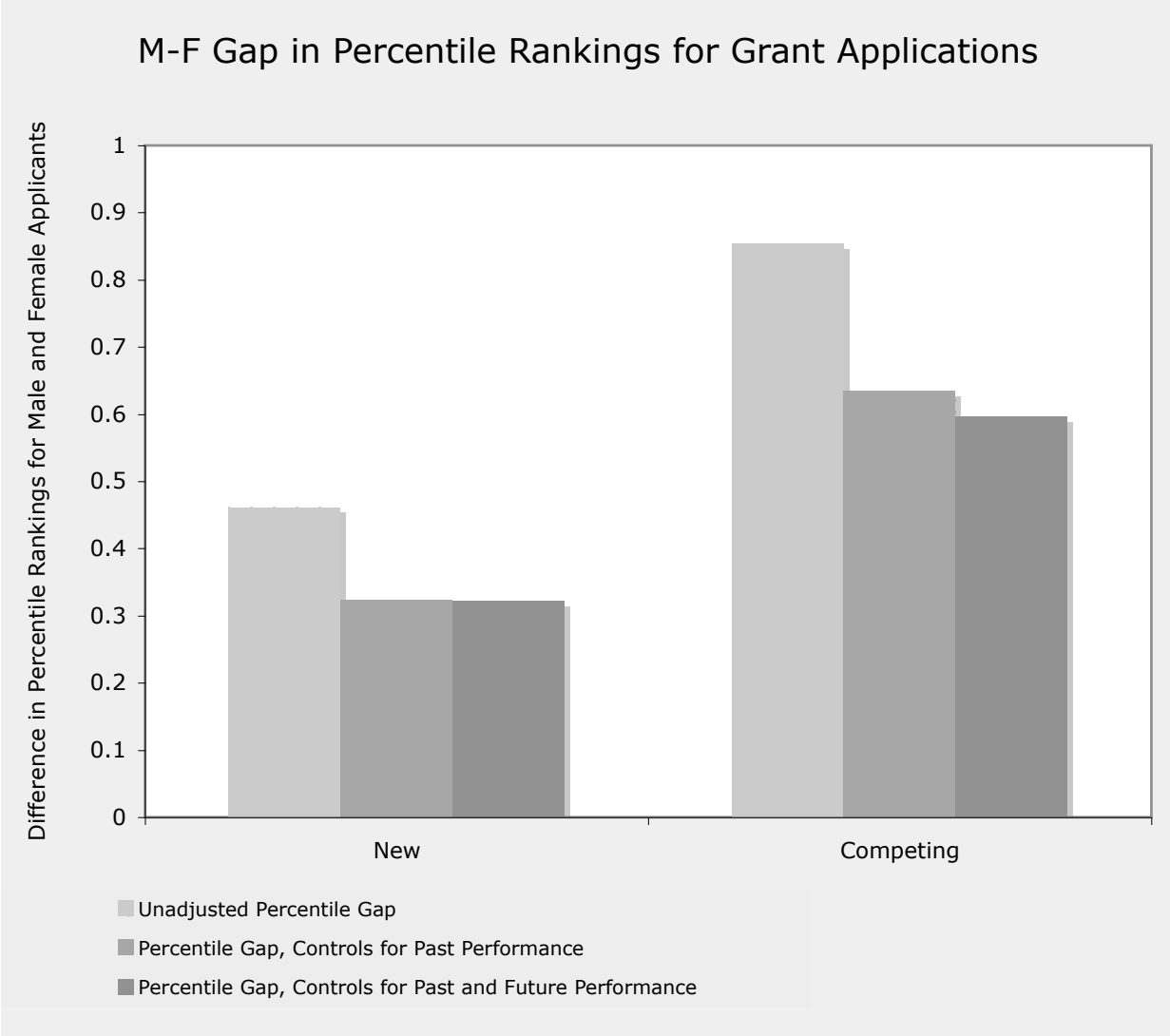

Figure 4: Female investigators receive lower percentile scores unconditionally, as well as conditionally on past and future performance. The level of bias is higher for competing grants.

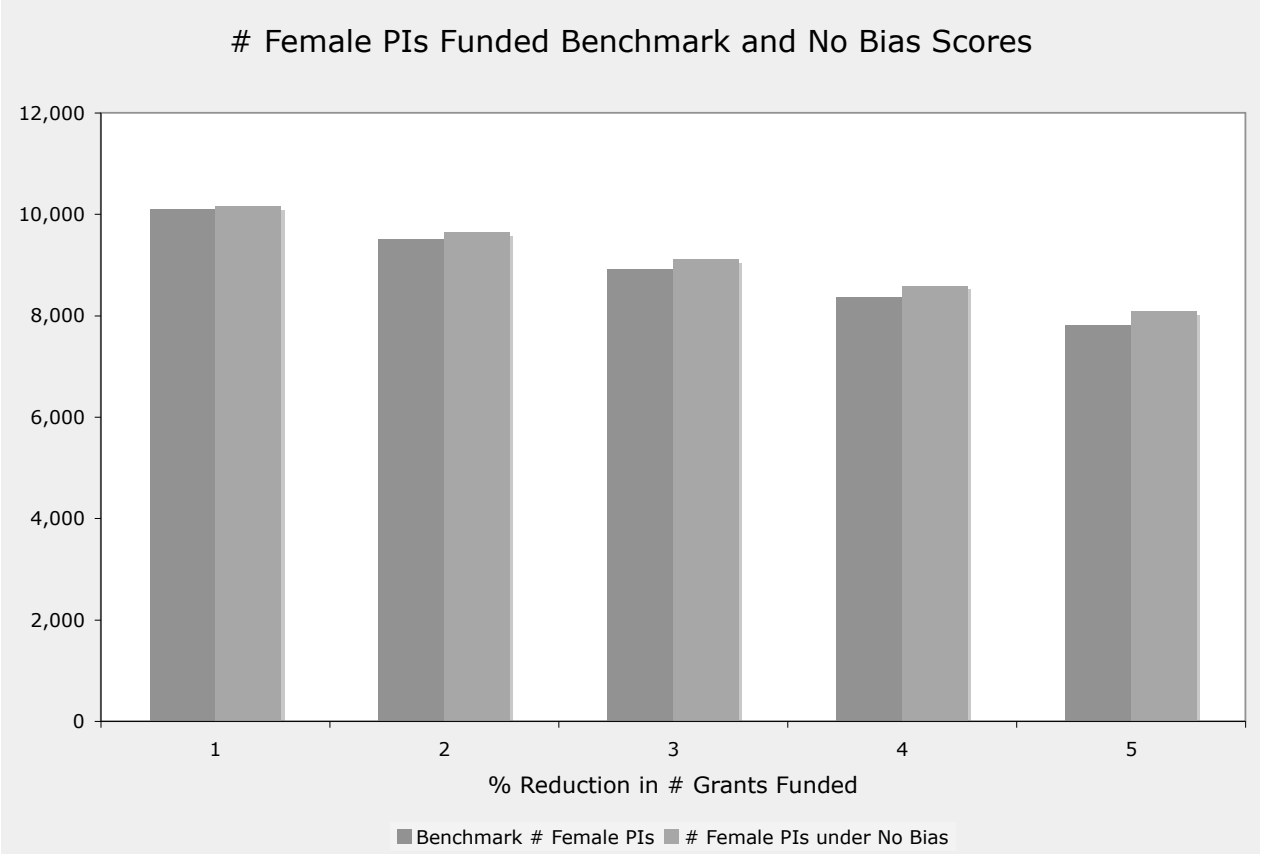

Figure 5: Fewer female investigators are funded as a result of taking gender into account.

A Context and Data

My data on R01 grants and their priority scores come from an NIH contract. Attendance rosters were collected for 286 chartered study sections from the NIH Center for Scientific Review. Data on publications and citations are linked to investigators via the Web of Science. Gender is defined probabilistically based on the first name of the PI and investigators are assumed to be female if the probability that they are female is greater than one half. Names for which gender probabilities could not be ascertained were dropped (5 percent of sample) and high frequency names were also dropped in order to ensure that matched publications were indeed the author's own (10 percent of sample).

B Methods

B.1 Identifying Bias

I use regression analysis to assess the extent of gender discrimination in NIH peer review. The raw gender gap in assigned percentiles is computed from the following regression:

$$R_{ist} = \alpha + \beta(\text{Applicant is Female}) + \delta_{st} + \varepsilon_{ist}. \tag{1}$$

Here, the percentile ranking R_{ist} received by applicant i to study section s at time t is modeled as a function of the applicant's gender. Fixed effects δ_{st} capture any unobserved differences in how individual study section meetings score grants so that β can be interpreted as the average difference in percentile ranking received by female and male applicants who were reviewed in the same meeting of the same study section.

To account for differences in qualifications among male and female applicants, I modify Equation (1) to include a set X_{ist} of variables describing the applicant's publication history, institutional affiliation, age, and prior grant history:

$$R_{ist} = \alpha + \beta(\text{Applicant is Female}) + \mu X_{ist} + \delta_{st} + \varepsilon_{ist}. \tag{2}$$

Specifically, X_{ist} includes controls for 1) the total number of citations that the PI received for all first, second, or last authored publications published in the five years prior to receiving the

grant, 2) the total number of first, second, or last authored publications in the 99.9, 99.5, 99, 95, 90, 75, 66, 50, and 25th percentiles of the citation distribution in the five years prior to receiving the grant, 3) dummy variables for the number of past successful new and competing R01s and other NIH grants, 4) career age dummies, 5) degree type dummies, and 6) fixed effects for a PI's institution FEs control for an applicant's institution at the time of application.

Given these controls, the coefficient β on the applicant's gender is interpreted as the percentile difference in scores between female and male applicants reviewed by the same study section meeting, who have similar past publications, grant histories, and who are affiliated with the same institution. Finally, in order to identify the portion of the percentile gap that is attributable to discrimination, I include additional controls P_{ist} for the future performance of the grant.

$$R_{ist} = \alpha + \beta_1(\text{Applicant is Female}) + \beta_2 P_{ist} + \mu X_{ist} + \delta_{st} + \varepsilon_{ist}. \tag{3}$$

The set of grant performance measures I use are: 1) the total number of citations that the PI received for all last authored publications published in the five years after receiving the grant and 2) the total number of last authored publications in the 99.9, 99.5, 99, 95, 90, 75, 66, 50, and 25th percentiles of the citation distribution in the five years prior to receiving the grant. In order to match publications to grants for earlier years, I restrict to PIs who have only one outstanding NIH grant at a time. Now β_1 measures the role of bias.

B.2 How Large Are These Effects?

To assess the consequences of taking gender into account for the number of female investigators who are funded, I construct a counterfactual portfolio of funded grants under the assumption that female investigators are treated the same as male investigators. This rules out both bias as well as different levels of stringency in review. To do this, I generate a hypothetical payline $\bar{R}^0$ such that anywhere from 5 to 25 percent of the grants I observe are cut. Using this new threshold, I calculate benchmark total number of funded women as the number of female PIs for all grants that fall below $\bar{R}^0$ according to their actual percentile rankings. I then generate counterfactual percentiles for each

of the cases above based on the estimated coefficients from Equation (2):

$$\begin{aligned} R_{ist}^{\text{Benchmark}} &= \widehat{\alpha} + \widehat{\beta}(\text{Applicant is Female}) + \widehat{\mu} X_{ist} + \widehat{\delta}_{st} \\ R_{ist}^{\text{Gender Neutral}} &= \widehat{\alpha} + \widehat{\mu} X_{ist} + \widehat{\delta}_{st} \end{aligned}$$

I rerank grant applications according to its counterfactual score and again consider the number of women, total number of citations, and total number of high impact publications for grants falling above the threshold according to both $R_{ist}^{\text{Benchmark}}$ and $R_{ist}^{\text{Gender Neutral}}$.

B.3 How Can Study Section Performance Be Improved?

To assess the impact of female study section membership on bias, I estimate the following regression model:

$$\begin{aligned} R_{ist} &= \alpha + \beta_1(\text{Applicant is Female}) + \beta_2(\text{Percentage Reviewers Female}) \\ &+ \beta_3(\text{Applicant is Female}) \times (\text{Percentage Reviewers Female}) \\ &+ \gamma Q_{ist} + \mu X_{ist} + \delta_{st} + \varepsilon_{ist}. \end{aligned}$$

Here, β_1 is the percentile gender gap for study sections with no female reviewers and β_3 identifies the change in the gender gap when the percentage of female reviewers increases.

TABLE 1: FUNDED GRANTS SUMMARY STATISTICS

	Full Sample		Roster Matched Sample	
		SD		SD
Sample Coverage				
# Grants	46,093		21,703	
# Applicants	24,257		15,128	
Years	1992-2007		1993-2005	
# Study Sections	495		286	
# Study Section Meetings	5,918		2,290	
Grant Characteristics				
% New Grants	55.49		54.71	
% Accepted at First Submission	49.87		49.97	
1-Percentile Priority Score (higher is better)	87.16	8.95	87.08	8.65
# Publications in 99th %tile of citations, next 5 years	1.13	2.63	1.63	3.16
# Publications in 95th %tile of citations, next 5 years	2.20	4.17	2.96	4.76
# Total Citations (100s), next 5 years	1.75	4.00	1.91	4.19
PI Characteristics				
% Female	23.26		23.43	
Years since last degree	19.55	9.12	19.64	9.19
% M.D.	27.08		26.85	
% Ph.D.	80.97		81.72	
# Past New or Competing R01s	3.45	3.47	3.51	3.51
# Publications in 99th %tile of citations, past 5 years	0.98	2.58	0.88	2.06
# Publications in 95th %tile of citations, past 5 years	2.76	5.02	2.77	4.47
# Total Citations (100s), past 5 years	4.79	7.18	5.09	7.48
Study Section Characteristics				
# Reviewers			20,847	
% Female			29.72	10.42
# Funded Grants			11.43	4.49

Notes: The full sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2007. The roster matched sample is a subsample that can be matched to the precise meeting (committee and date) in which they were scored. Future Publications refers to the number of last-authored publications that the grant winner publishes in the 5 years following the grant which fall into the top X-percentile of the citation distribution for publications published in the same year. Past publications is defined analogously, except that I include first and second authored articles in addition to last authored articles.

TABLE 2: CORRELATIONS BETWEEN APPLICANT GENDER AND PERCENTILE RANKING

	(1)	(2)	(3)	(4)	(5)	(6)
			Controls for Past Performance		Controls for Past and Future Performance	
	New	Competing	New	Competing	New	Competing
Female	-0.461***	-0.854***	-0.324*	-0.634***	-0.324*	-0.634***
	(0.00556)	*(1.09e-05)*	*(0.0547)*	*(0.00123)*	*(0.0547)*	*(0.00123)*
Observations	25576	20517	25576	20517	25576	20517
R-squared	0.324	0.294	0.338	0.323	0.338	0.323
Meeting FE	X	X	X	X	X	X
Past Performance, Demographics, Institution FEs			X	X	X	X
Future Performance					X	X

Notes: Sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2007. PIs are restricted to those with unique first and last name combinations in the PubMed database. Past performance includes controls for the number of citations for publications published five years prior to receiving the grant, and the number of publications in the 99.9, 99.5, 99, 95, 90, 75, 66, 50, and 25th percentiles of the citation distribution in the five years prior to receiving the grant. Past performance also includes controls for the number of past successful new and competing R01s and other NIH grants. Demographics include controls for career age dummies, and type of degree. Insitution FEs control for an applicant's institution at the time of application, where available. Controls for future performance includes controls for the number of citations for publications published 5 years subsequent to receiving the grant, and the number of publications in the 99.9, 99.5, 99, 95, 90, 75, 66, 50, and 25th percentiles of the citation distribution in the five years subsequent to receiving the grant.

TABLE 3: HOW WOULD THE NIH GRANT PORTFOLIO LOOK IF PEER REVIEW WERE GENDER NEUTRAL?

	Total # of Female PIs Funded	
	Benchmark	Gender Neutral
Top 95% Funded		
	10,104	10,154
(% change relative to benchmark)		*0.49*
Top 90% Funded		
	9,500	9,645
(% change relative to benchmark)		*1.53*
Top 85% Funded		
	8,923	9,110
(% change relative to benchmark)		*2.10*
Top 80% Funded		
	8,355	8,586
(% change relative to benchmark)		*2.76*
Top 75% Funded		
	7,813	8,082
(% change relative to benchmark)		*3.44*

Notes: These are calculated for counterfactual funding thresholds as described in the text. The benchmark is given by fitted scores, not actual scores.

TABLE 4: EFFECT OF STUDY SECTION COMPOSITION ON SCORING

	(1)	(2)	(3)	(4)
			Controls for % Female on Study Section	
	New	Competing	New	Competing
Female	0.060	-0.367	-0.216	-2.025**
	(0.798)	*(0.199)*	*(0.710)*	*(0.0149)*
Female X (% Study Section is Female)			0.967	6.006**
			(0.607)	*(0.0352)*
Observations	11874	9829	11874	9829
R-squared	0.293	0.294	0.293	0.294
Meeting FE	X	X	X	X
Past Performance, Demographics, Institution FEs	X	X	X	X
Detailed Controls for Future Performance.	X	X	X	X

Notes: Sample includes new or competing R01 grants evaluated in charterd study sections from 1992 to 2007. PIs are restricted to those with unique first and last name combinations in the PubMed database. Past performance includes controls for the number of citations for publications published five years prior to receiving the grant, and the number of publications in the 99.9, 99.5, 99, 95, 90, 75, 66, 50, and 25th percentiles of the citation distribution in the five years prior to receiving the grant. Past performance also includes controls for the number of past successful new and competing R01s and other NIH grants. Demographics include controls for career age dummies, and type of degree. Insitution FEs control for an applicant's institution at the time of application, where available. Detailed controls for future performance includes controls for the number of citations for publications published 5 years subsequent to receiving the grant, and the number of publications in the 99.9, 99.5, 99, 95, 90, 75, 66, 50, and 25th percentiles of the citation distribution in the five years subsequent to receiving the grant.