Female grant applicants are equally successful when peer reviewers assess the science, but not when they assess the scientist

Holly O. Witteman, PhD[1]; Michael Hendricks, PhD[2];

Sharon Straus, MD FRCPC MSc HBSc[3]; Cara Tannenbaum, MD MSc[4]

[1]Associate Professor, Department of Family and Emergency Medicine, Faculty of Medicine, 1050 avenue de la Médecine, Université Laval, Quebec City, Canada, G1V 0A6

[2]Assistant Professor, Department of Biology, Faculty of Science, 1205 av du Docteur-Penfield, McGill University, Montreal, Canada, H3A 1B1

[3]Professor, Li Ka Shing Knowledge Institute of St. Michael's Hospital, 30 Bond Street, Toronto, Canada, M3B 2T9

[4]Scientific Director, Institute for Gender and Health, Canadian Institutes of Health Research, Ottawa, Canada, H3A 1W4

Corresponding author: Holly O. Witteman, 1050 avenue de la Médecine, Pavillon Ferdinand-Vandry, Université Laval, Québec City, Québec, Canada, G1V 0A6, telephone: 418.656.2131 x3981, fax: 418.656.2465, email: holly.witteman@fmed.ulaval.ca

Keywords (MeSH terms): Peer Review, Research/methods; Sex Factors

Keywords (non-MeSH terms): Gender Bias

Running title: Gender Bias in Grant Review

Word count: 3500 words

ABSTRACT (299 words)

Background: Previous research shows that men often receive more research funding than women, but does not provide empirical evidence as to why this occurs. In 2014, the Canadian Institutes of Health Research (CIHR) created a natural experiment by dividing all investigator-initiated funding into two new grant programs: one with and one without an explicit review focus on the caliber of the principal investigator.

Methods: We analyzed application success among 23,918 grant applications from 7,093 unique principal investigators in a 5-year natural experiment across all investigator-initiated CIHR grant programs in 2011-2016. We used Generalized Estimating Equations to account for multiple applications by the same applicant and an interaction term between each principal investigator's self-reported sex and grant programs to compare success rates between male and female applicants under different review criteria.

Results: The overall grant success rate across all competitions was 15.8%. After adjusting for age and research domain, the predicted probability of funding success in traditional programs was 0.9 percentage points higher for male than for female principal investigators (OR 0.934, 95% CI 0.854-1.022). In the new program focused on the proposed science, the gap was 0.9 percentage points in favour of male principal investigators (OR 0.998, 95% CI 0.794-1.229). In the new program with an explicit review focus on the caliber of the principal investigator, the gap was 4.0 percentage points in favour of male principal investigators (OR 0.705, 95% CI 0.519-0.960).

Interpretation: This study suggests gender gaps in grant funding are attributable to less favourable assessments of women as principal investigators, not differences in assessments of the quality of science led by women. We propose ways for funders to avoid allowing gender bias to influence research funding.

Funding: This study was unfunded.

INTRODUCTION

For decades, studies have shown that women in academia and science must perform to a higher standard than men to receive equivalent recognition,[1–4] especially Indigenous and racialized women.[5–11] Compared to men, women are more often characterized as lacking the brilliance, drive, and talent required to carry a novel line of inquiry through to discovery,[8,12] with children as young as six years old endorsing such stereotypes.[13] Women are less likely than men to be viewed as scientific leaders[14–17] or depicted as scientists.[18] Women in academia contribute more labor for less credit on publications,[19,20] receive less compelling letters of recommendation,[21–24] receive systematically lower teaching evaluations despite no differences in teaching effectiveness,[25] are more likely to experience harassment,[11,26–29] and are expected to do more service work[30,31] and special favors for students.[32] While men in academia have more successful careers after taking parental leave, women's careers suffer after the same.[33] Women receive less start-up funding as biomedical scientists[34] and are underrepresented in invitations to referee papers.[35] Compared to publications led by men, those led by women take longer to publish[36] and are cited less often,[37,38] even when published in higher-impact journals.[39] Papers[40] and conference abstracts[41] led by women are accepted more frequently when reviewers are blinded to the identities of the authors. Women are underrepresented as invited speakers at major conferences,[4] prestigious universities,[42] and grand rounds.[43,44] When women are invited to give these prestigious talks, they are less likely to be introduced with their formal title of Doctor.[45] Female surgeons have been shown to have better patient outcomes overall,[46] yet, when a patient dies in surgery under the care of a female surgeon, general practitioners reduce referrals to her and to other female surgeons in her specialty, whereas they show no such reduced referrals to male surgeons following a patient's death.[47] Women are less likely to reach higher ranks in medical schools even after accounting for age, experience, specialty, and measures of research productivity.[48,49] When fictitious or real people are presented as women in randomized experiments, they receive lower ratings of competence from scientists,[50] worse teaching evaluations from students,[51] and fewer email responses from professors after presenting as students seeking a PhD advisor[9] or as scientists seeking copies of a paper or data for a meta-analysis.[52]

Conversely, other research has demonstrated advantages experienced by women in academia; for example, achieving tenure with fewer publications than men.[53] In assessments of potential secondary and postsecondary teachers and professors, women are favored in male-dominated fields, as are men in female-dominated fields.[54] When fictitious people are presented as women

in randomized experiments, they receive higher rankings as potential science faculty.[55,56] This aligns with evidence from other contexts showing that high-potential women are favored over high-potential men[57] and that, while women face discrimination at earlier stages, once women have proven themselves in a male-dominated context, they are favored over men.[58]

In sum, there is considerable evidence that women face or have faced persistent, pervasive barriers in many aspects of academia and science. There is also recent evidence that women in academia and science may fare better in merit evaluations than equally-qualified men after they have progressed beyond postsecondary education. In light of this evidence as a whole, we consider the question: does bias in favor of men or women influence research grant funding?

Previous research on this question has been suggestive but not conclusive. A 2007 meta-analysis of 21 studies from a range of countries found an overall gender gap in favor of men, with 7% higher odds of fellowship or grant funding for male applicants.[59] Research since then has documented that, compared to their male colleagues, female principal investigators have lower grant success rates,[60] lower grant success rates in some but not all programs,[61,62] equivalent grant success rates after adjusting for academic rank[63,64] but fewer funds requested and received,[63–66] or equivalent funding rates.[67] To the best of our knowledge, no such study has yet found women to experience higher grant success rates nor to receive more grant funding than men.

These previous studies have been observational, making it difficult to draw robust conclusions about the causes of sex-related discrepancies in funding success rates when they are observed. Furthermore, some previous studies have not accounted for potential confounding variables; for example, domain of research.[59,68,69] Our objective in this study was therefore to determine whether gender gaps in grant funding are attributable to potential differences in how male and female principal investigators are evaluated, using real-world data and a study design that would allow for stronger conclusions than those from observational studies. Our study made use of a natural experiment at a national health research funding agency which allowed for the comparison of grant success rates among male and female principal investigators between three grant programs. The three programs were: traditional grant programs, which had demonstrated higher success rates among younger male principal investigators for applications submitted in 2001-2011[60] but for which no such analysis had been conducted after 2011, and

two new competitions, one with and without an explicit review focus on the quality of the principal investigator.

Our hypotheses regarding comparisons of gaps in grant success rates between female and male principal investigators after accounting for age and research domain were as follows.

> H_0: Gaps will be similar under traditional review criteria and both sets of new review criteria.
> H_1: The gap will be larger in favour of male principal investigators in the new competition with more focus on the science.
> H_2: The gap will be larger in favour of male principal investigators in the new competition with more focus on the scientist.

Support for H_0 would suggest that gaps, when present, may reflect different career paths and choices made by women and men,[70,71] differences between the types of research proposed by female and male principal investigators, or may be spurious. Support for H_1 would suggest that gaps are are partly or wholly driven by female principal investigators proposing science assessed as lower quality than that of their male colleagues. Support for H_2 would suggest that gaps are partly or wholly driven by women being assessed less favorably as principal investigators compared to their male colleagues. Other potential alternative hypotheses such as gaps in favour of female principal investigators were not considered a priori because publicly-available summary statistics of the programs showed such results to be impossible.

METHODS

Population, Setting and Data Source

As depicted in Figure 1, beginning in 2014, the Canadian Institutes of Health Research (CIHR) phased out traditional open grant programs, dividing investigator-initiated funding into two new open programs: the Project grant program and Foundation grant program. Both new programs used a staged review process in which lower-ranked applications are rejected from continuing on to further stages. As in traditional programs, in the new Project grant program, reviewers were instructed to primarily assess the research proposed. In contrast, the Foundation grant program was about ‘people, not projects’. Reviewers at the first stage were instructed to primarily assess the principal investigator, with 75% of the score being allocated to reviewers’ assessments of the applicant’s leadership, productivity, and the significance of their

contributions.[72] Table 1 lists the adjudication criteria for each of these. Only applicants who passed this stage were invited to submit a detailed proposal describing their research.

Analysis

We analyzed data from all applications submitted to CIHR grant programs across all investigator-initiated competitions from 2011 through 2016. We used Generalized Estimating Equations to fit a logistic model that accounted for applicants submitting multiple applications during these five years.[73,74] We modeled grant success rates as a function of the grant program, applicants' self-reported binary sex, self-reported age, self-declared domain of research, and an interaction term between applicant sex and grant program. We controlled for age and domain of research because younger cohorts of investigators include larger proportions of women, as do domains of health research other than biomedical. Full methodological details are available in the online appendix.

RESULTS

The data set contained 23,918 applications from 7,093 unique applicants. There were 15,775 applications from 4,472 male applicants and 8,143 applications from 2,621 female applicants. The overall grant success rate across the data set was 15.8%. As shown in Figure 2, after adjusting for age and research domain, the predicted probability of funding was 0.9 percentage points higher for male applicants than female applicants in traditional grant programs (OR 0.934, 95% CI 0.854-1.022), 0.9 percentage points higher in the Project program (OR 0.998, 95% CI 0.794-1.229) and 4.0 percentage points higher in the Foundation program (OR 0.705, 95% CI 0.519-0.960). Figure 3 shows how the gap in the Foundation program was driven by discrepancies at the first review stage, where review focused on the principal investigator. Odds of receiving funding were also lower in the three non-biomedical domains and higher for older applicants. Tabular results are available in the online appendix.

DISCUSSION

Our study provides stronger evidence than was previously available regarding the likely causes of gender gaps in grant funding. In this natural experiment, when reviewers primarily assessed the science, there were no statistically significant differences between success rates for male and female principal investigators. When reviewers explicitly assessed the principal investigator as a scientist, the gap was significantly larger. These data support the second of our alternative

hypotheses; namely, that gender gaps in funding stem from female principal investigators being evaluated less favourably than male principal investigators, not from differences in evaluations of the quality of their science.

The finding that women are evaluated less favourably than men as applicants aligns with previous studies from other countries. Data from the United States showed that female grant applicants to the National Institutes of Health's flagship R01 program were less likely than male applicants to be described as leaders.[14] In the Netherlands, grant reviewers gave equal scores to men's and women's proposed research but assigned lower scores to women as researchers.[75] In Sweden, similar differences have been shown among evaluators' assessments of applicants for governmental venture capital.[76]

There are two main reasons that female principal investigators might be evaluated less favourably than male principal investigators. First, reviewers may exhibit gender-biased subjective evaluations of principal investigators. Second, female principal investigators may have submitted weaker applications than male peers. Below, we consider each of these two potential reasons in the context of available literature.

The suggestion that reviewers might exhibit conscious or unconscious bias against women aligns with previous evidence about gender gaps at the highest levels of achievement. For example, gender gaps in top mathematics performers have been shown to be attributable to cultural gender inequality, not differences in natural ability.[77] In orchestra auditions, women became more successful when auditioning musicians' identities were concealed behind a screen.[78] Similar increases in success have occurred for women submitting manuscripts[40] and conference abstracts[41] after reviewers were blinded. Randomized experiments using hypothetical scenarios (e.g., "How would you respond if …") have shown scientists to favour hypothetical men as laboratory managers,[50] hypothetical men as junior faculty in an experiment conducted in 1997-1998,[79] but hypothetical women as junior faculty in a more recent experiment.[55,56] Randomized experiments in which the outcome was real behaviour all favoured men,[51,52] particularly white men.[9]

If Foundation Grant reviewers' evaluations of applicants were gender-biased, it is possible that such bias might have been due to elements outside reviewers' control; for example, the structure of the application. The Foundation Grant Stage 1 application form was accompanied

by a structured curriculum vitae (CV) describing the last seven years of the principal investigator's research. During this study, the seven-year limit was not extended in the case of a maternity, parental, or other leave. It is possible that the inability to list a full seven years' worth of productivity in the case of leaves may have more negatively influenced reviewers' evaluations of female compared to male applicants. Female principal investigators may also have described their accomplishments more modestly, though, to our knowledge, there are no data suggesting this gendered stereotype applies to competitive scientists.

The idea that reviewers' evaluations might reflect weaker applications by female principal investigators requires that female principal investigators showed weaker leadership, less significant contributions, lower productivity, or less compelling visions and program directions compared to male principal investigators. Regarding leadership, previous evidence shows women are less likely to advance to leadership roles in medicine, even after accounting for confounding variables such as differences in publication records.[49,80] Women's CVs are less likely to include prestigious invited talks.[4,42–44] The leadership credentials of women in health research may also be constrained by broader societal tendencies in which women receive negative responses for being ambitious[81] or exerting control as leaders.[82–84] Research suggests that women must choose between being seen as a competent leader or being liked,[82,83] which complicates career advancement within academia, where leadership positions often depend on colleagues' endorsements. Societal stereotypes suggest that women are less interested in leadership positions, but to our knowledge there is no evidence that this is true for women in health research.

Regarding significance of contributions, papers led by women have been shown to be cited less often,[37,38] which may indicate lower significance of contributions. It is possible that women's track records of impactful work may be negatively affected by expectations of service and other work[30–32] or by the "productivity tax"[85] of dealing with harassment.[11,26–28] Data are lacking about men and women's records regarding launching the careers of trainees, but research suggests that trainees publish more under the supervision of female principal investigators.[86,87]

Regarding productivity, globally, there are more publications authored by men.[37,88] Ignoring the potential confounder of different publication norms in fields with different gender ratios, papers from Canada had a somewhat lower proportion of female fractional authorships (31% for publications in 2008-2012[89]) than national statistics of female faculty might suggest (35% at

rank assistant professor or higher in 2010-2011[90]). Male principal investigators in our study may also have had more publications due to historically-higher CIHR grant funding yielding more data for them to publish.[60] Together, this evidence suggests that it is possible that female principal investigators in our study had fewer publications compared to their male peers, though this may be confounded by career stage. Research by Tamblyn and colleagues is examining male and female principal investigators' productivity records within the Foundation grant program and whether these were evaluated equivalently by reviewers. Older research showed that reviewers in Sweden assessed equivalent productivity less positively for female fellowship applicants compared to male applicants,[2] but this study analyzed reviews conducted nearly twenty-five years ago and may not apply in our context.

Finally, regarding vision and program direction, to the best of our knowledge, there is no previous evidence about gender gaps or lack thereof in evaluations of such brief descriptions of research programs. The lack of difference in evaluations of Project Grants suggests that it is unlikely there were inherent differences in the quality of program descriptions.

In summary, we cannot pinpoint whether female principal investigators may have been evaluated more negatively due to bias on the part of reviewers or to review criteria designed around metrics that may reflect lower performance or cumulative disadvantage. Previous literature suggests that both individual and structural bias are plausible. There is more experimental evidence supporting the potential for individual bias than structural bias or objective differences in performance, but it is arguably harder to run randomized experiments about structural bias. Perhaps most importantly, from a policy perspective, it matters little which of these is the most likely cause. If, indeed, there are systematic differences between the leadership, contribution and productivity metrics of male and female principal investigators, is a funding competition based on these metrics justified when there is no indication that women's science is of lower quality?

The CIHR deemed such imbalanced outcomes not justifiable. Following the grant cycles analyzed in our study, and as part of a broader Equity Framework,[91] the CIHR implemented new policies to eliminate the observed gender gap in Foundation grants. The policies included instructions that reviewers should complete an evidence-based[92] reviewer training module about multiple forms of unconscious bias.[93] Training has been shown to help mitigate the effects of bias in some studies[15,94] though not in others.[95,96] The Foundation grant program regulations

were also revised such that, should reviewer training not have the desired effects at Stage 1, a proportional number of female applicants would nonetheless proceed to the next stage at which their proposed research would be evaluated. In other contexts, such quotas based on the available pool of candidates have been shown to increase women's representation, the overall quality of candidates, or both.[97–99] In the following Foundation grant cycle, which was underway during this project, success rates were equivalent for male and female principal investigators and the policy did not need to be invoked. (See summary data in the online appendix.) In the subsequent cycle, the reviewer training module alone did not result in equivalent success rates at Stage 1 among male and female principal investigators, and the policy of equivalence was invoked.[100]

Our study had three main limitations. First, principal investigators were not randomized to one grant program or the other. Although aspects of our study minimized the potential to observe the results we found, the non-randomized design leaves open the possibility that unobserved confounders or selection bias may have contributed to observed differences. Second, our planned analyses did not include investigation of 3-way analyses between grant program group, self-reported sex, and age. It is possible that the overall two-way interaction observed might be driven by differences in evaluation of men and women who are more advanced in their careers. Our planned analyses did not include this, as we were concerned about insufficient sample size to include three-way interactions in the model. Third, we assumed that people who self-report as female or male also identify as women or men, respectively. This may not be true in all cases. Data are lacking regarding how many people identify as transgender or non-binary in Canada. Context may be offered by a recent analysis suggesting that transgender people were 0.4% of the US population.[101] If this proportion were reflected in our study and if all transgender applicants reported their sex as their assigned sex at birth, this would represent 28 people in total within this study, a number unlikely to meaningfully change the results of our analyses.

Our study also has four main strengths. First, it was quasi-experimental. To the best of our knowledge, this is the first evidence from a study design that was not fully observational. This enables stronger conclusions than were previously possible. Second, while quasi-experimental studies have potential for selection bias,[102] in this study, selection bias was limited by eligibility rules. Specifically, for principal investigators who had already established their careers and received funding from the CIHR, only those whose grants from traditional programs were expiring within a specific time period were permitted to apply to Foundation. This meant that a

portion of allocation was dependent on an external, arbitrary variable. Third, we controlled for age and domain of research, two key confounders in studies of gender bias in grant funding. Academic rank, which correlates with age, has been shown to account for gender gaps in grant funding in other studies, as has domain of research.[68,69] Having accounted for these key confounders strengthens our findings. Fourth and finally, our study analyzed all available data over a period of five years from a major national funding agency, thus offering evidence from a large data set of real-world grant review.

Conclusions

Bias in grant review prevents the best research from being funded. When this occurs, lines of research go unstudied, careers are damaged, and funding agencies are unable to deliver the best value for money, not only within a given funding cycle, but also long term as small differences compound into cumulative disadvantage. 'People, not projects' funding programs may be at higher risk of reproducing and exacerbating common societal biases in research funding. To encourage rigorous, fair peer review that results in funding the best research, we recommend that funders focus assessment on the science rather than the scientist, measure and report funding by applicant characteristics and potential confounding variables, and consider reviewer training and other policies to mitigate the effects of all forms of bias. Future research should investigate the role of confounding variables not included in our study, other types of bias beyond gender, and methods of increasing fairness and rigour in peer review.

FIGURES

Figure 1. Changes in grant programs

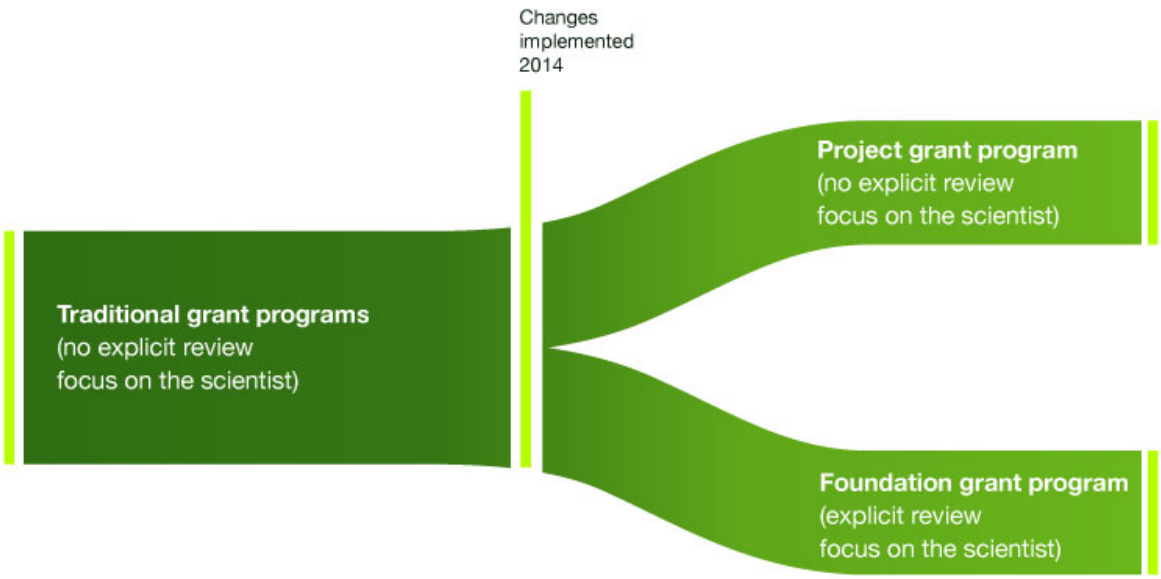
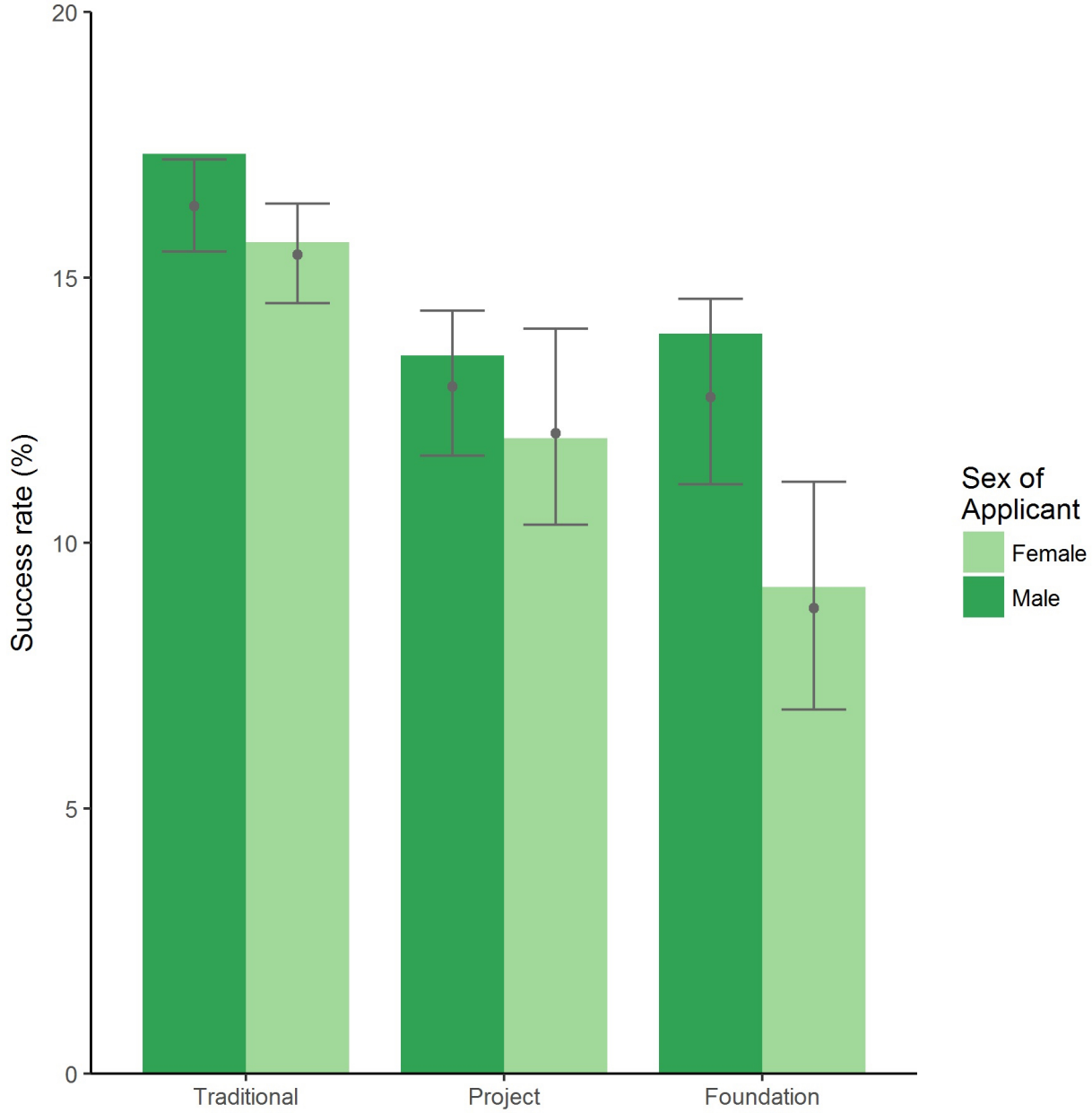
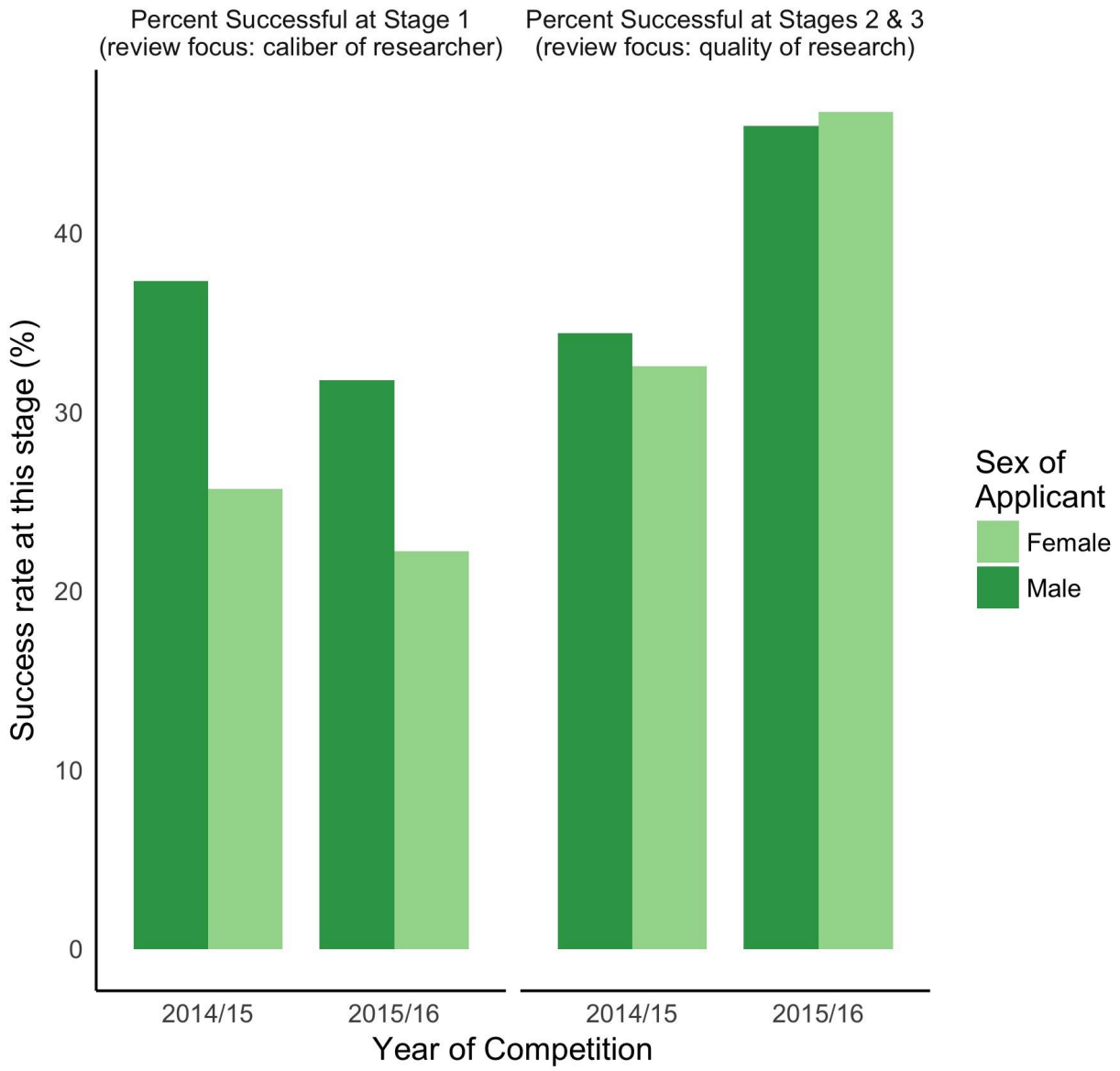


Figure 2. Funding success rate by grant program



Columns indicate observed success rates. Points and error bars indicate model-predicted means and 95% confidence intervals, respectively.

Figure 3. Foundation results by review stage



Columns indicate observed success rates.

TABLES

Table 1. Foundation Grant Stage 1 Application Adjudication Criteria

Application Section	Maximum Length	Score Weight	Criteria
Leadership	1750 characters, including spaces (about half a page)	25%	• Is the applicant recognized in their field, demonstrating a history of holding influential roles, inspiring others, mobilizing communities and advancing the direction of a field? • Has the applicant demonstrated the ability to successfully establish, resource, and direct programs of research, which should include: securing the required resources, ensuring effective collaboration, and/or incorporating knowledge translation strategies?
Significance of Contributions	1750 characters, including spaces (about half a page)	25%	• Has the applicant significantly advanced knowledge and/or its translation into improved health care, health systems, and/or health outcomes? • Has the applicant engaged, trained, and/or launched the career paths of promising individuals in research and/or other health-related non-academic fields?
Productivity	1750 characters, including spaces (about half a page)	25%	• Has the applicant demonstrated an outstanding level (quantity) of research outputs in their field based on prior work? • Has the applicant's previous work generated high quality research outputs?
Vision and Program Direction	3500 characters, including spaces (about a page)	25%	• Are the vision, goal, overall objective(s), and potential contribution(s) of the proposed research program well-defined and well-articulated in the context of a logical career progression for the Program Leader(s)? • Is the vision forward-looking, creative, and appropriately ambitious? • Does the vision aim to significantly advance knowledge and/or its translation to improved health care, health systems, and/or health outcomes?

APPENDICES

1. Methodological and Analytical Details
2. R Code

DECLARATIONS

Abbreviations

CIHR: Canadian Institutes of Health Research

Ethics Approval, Consent to Participate and Consent for Publication

The views expressed in this paper are those of the authors and do not necessarily reflect those of the CIHR or the Government of Canada. Data were held internally and analyzed by staff at the CIHR within their mandate as a national funding agency. Research and analytical studies at the CIHR fall under the Canadian Tri-Council Policy Statement 2: Ethical Conduct for Research Involving Humans (available: pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/Default/, accessed 2017 July 13.) This study had the objective of evaluating CIHR's Investigator-Initiated programs, and thus fell under Article 2.5 of TCPS-2 and not within the scope of Research Ethics Board review in Canada. Nevertheless, applicants were informed through ResearchNet, in advance of peer review, that CIHR would be evaluating its own processes. All applicants provided their electronic consent; no applicant refused to provide consent.

Availability of Data and Materials

Data are confidential due to Canadian privacy legislation. Researchers interested in addressing other research questions related to grant funding may contact the CIHR at funding-analytics@cihr.ca.

Competing Interests

This work was unfunded. HW holds grant funding from the CIHR as principal investigator of a Foundation grant. HW and MH are two of the three founding national co-chairs of the Association of Canadian Early Career Health Researchers, an organization that has published statements critical of aspects of the CIHR grant program changes, including the Foundation grant program. CT is a Scientific Institute director at the CIHR and is therefore partially employed by the CIHR. SS holds grant funding from the CIHR, including a Foundation grant, and also received contract funding from the CIHR to lead the scoping review described herein and to analyze applicant and reviewer survey responses. HW receives salary support from a Research Scholar Junior 1 Career Development Award from the Fonds de Recherche du

Québec-Santé. SS receives salary support from a Tier 1 Canada Research Chair in Knowledge Translation and Quality of Care.

Contributions

Conceptualization: HW MH DG (see Acknowledgements); Methodology: HW JW RD AC MH DG; Formal Analysis: JW; Investigation: HW JW RD AC; Data Curation: JW RD AC; Writing – Original Draft: HW JW; Writing – Review & Editing: HW JW RD AC MH CT SS DG; Visualization: HW JW RD; Project Administration: HW RD. Contributions are listed according to the CReDIT taxonomy (docs.casrai.org/CRediT).

Acknowledgements

The authors gratefully acknowledge early work by Ms. Anne-Sophie Julien, statistician, who worked with HW to outline potential analytical approaches, drafted preliminary R code for the proposed research question, and reviewed a draft of the manuscript. We also thank Ms. Anne Strong, CIHR staff, for extracting the dataset, and Dr. Anne Martin-Matthews, CIHR Vice-President, for her comments on drafts of this manuscript, and many colleagues who commented online in response to previous preprint versions of this manuscript. The authors are extremely grateful to Dr. Jonathan A. Whiteley (JW), Dr. Rachelle E. Desrochers (RD), Dr. Alysha Croker (AC), and Dr. Danika Goosney (DG), all of whom were employed at the CIHR during this project and assisted with the scoping of the analysis and interpretation of the results. JW conducted the analysis at the CIHR as the data used consisted of confidential applicant information that could not be shared externally. He also drafted details of the methodology in collaboration with HW. The CIHR is committed to collaborating with the research community to advance knowledge on best practices in peer review without unduly influencing the conclusions drawn by its collaborators. For this reason, CIHR employees are not currently permitted to be co-authors on papers describing analyses of the agency's programs. Without the active participation of these CIHR employees and their contributions, this paper would not have been possible.

REFERENCES

1 Rossiter MW. The Matthew Matilda Effect in Science. *Soc Stud Sci* 1993; **23**: 325–41.

2 Wenneras C, Wold A. Nepotism and sexism in peer-review. *Nature* 1997; **387**: 341–3.

3 Reuben E, Sapienza P, Zingales L. How stereotypes impair women's careers in science. *Proc Natl Acad Sci U S A* 2014; **111**: 4403–8.

4 Klein RS, Voskuhl R, Segal BM, *et al.* Speaking out about gender imbalance in invited speakers improves diversity. *Nat Immunol* 2017; **18**: 475–8.

5 Settles IH, Cortina LM, Malley J, Stewart AJ. The Climate for Women in Academia: the Good, the Bad, and the Changeable. *Psychol Women Q* 2006; **30**: 47–58.

6 Ginther DK, Schaffer WT, Schnell J, *et al.* Race, ethnicity, and NIH research awards. *Science* 2011; **333**: 1015–9.

7 Williams J, Phillips KW, Hall EV. Double jeopardy?: Gender bias against women of color in science. Hastings College of the Law, Center for WorkLife Law, 2014.

8 Leslie S-J, Cimpian A, Meyer M, Freeland E. Expectations of brilliance underlie gender distributions across academic disciplines. *Science* 2015; **347**: 262–5.

9 Milkman KL, Akinola M, Chugh D. What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *J Appl Psychol* 2015; **100**: 1678–712.

10 Henry F, Dua E, Kobayashi A, *et al.* Race, racialization and Indigeneity in Canadian universities. *Race Ethnicity and Education* 2017; **20**: 300–14.

11 Clancy KBH, Lee KMN, Rodgers EM, Richey C. Double jeopardy in astronomy and planetary science: Women of color face greater risks of gendered and racial harassment. *J Geophys Res Planets* 2017; **122**: 2017JE005256.

12 Elmore KC, Luna-Lucero M. Light Bulbs or Seeds? How Metaphors for Ideas Influence Judgments About Genius. *Soc Psychol Personal Sci* 2016; published online Oct 7. DOI:10.1177/1948550616667611.

13 Bian L, Leslie S-J, Cimpian A. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science* 2017; **355**: 389–91.

14 Magua W, Zhu X, Bhattacharya A, *et al.* Are Female Applicants Disadvantaged in National Institutes of Health Peer Review? Combining Algorithmic Text Mining and Qualitative Methods to Detect Evaluative Differences in R01 Reviewers' Critiques. *J Womens Health* 2017; **26**: 560–70.

15 Carnes M, Devine PG, Baier Manwell L, *et al.* The effect of an intervention to break the gender bias habit for faculty at one institution: a cluster randomized, controlled trial. *Acad Med* 2015; **90**: 221–30.

16 Smyth FL, Nosek BA. On the gender-science stereotypes held by scientists: explicit accord

with gender-ratios, implicit accord with scientific identity. *Front Psychol* 2015; **6**: 415.

17 Carli LL, Alawa L, Lee Y, Zhao B, Kim E. Stereotypes about gender and science: Women ≠ scientists. *Psychol Women Q* 2016; **40**: 244–60.

18 Kerkhoven AH, Russo P, Land-Zandstra AM, Saxena A, Rodenburg FJ. Gender Stereotypes in Science Education Resources: A Visual Content Analysis. *PLoS One* 2016; **11**: e0165037.

19 Macaluso B, Larivière V, Sugimoto T, Sugimoto CR. Is Science Built on the Shoulders of Women? A Study of Gender Differences in Contributorship. *Acad Med* 2016; **91**: 1136–42.

20 Feldon DF, Peugh J, Maher MA, Roksa J, Tofel-Grehl C. Time-to-Credit Gender Inequities of First-Year PhD Students in the Biological Sciences. *CBE Life Sci Educ* 2017; **16**. DOI:10.1187/cbe.16-08-0237.

21 Trix F, Psenka C. Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty. *Discourse & Society* 2003; **14**: 191–220.

22 Schmader T, Whitehead J, Wysocki VH. A Linguistic Comparison of Letters of Recommendation for Male and Female Chemistry and Biochemistry Job Applicants. *Sex Roles* 2007; **57**: 509–14.

23 Madera JM, Hebl MR, Martin RC. Gender and letters of recommendation for academia: agentic and communal differences. *J Appl Psychol* 2009; **94**: 1591–9.

24 Dutt K, Pfaff DL, Bernstein AF, Dillard JS, Block CJ. Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nat Geosci* 2016; **9**: 805–8.

25 Stark P, Ottoboni K, Boring A, Cetinkaya-Rundel M. Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness. 2016; published online Jan 7. DOI:10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

26 Berdahl JL. The sexual harassment of uppity women. *J Appl Psychol* 2007; **92**: 425–37.

27 Jagsi R, Griffith KA, Jones R, Perumalswami CR, Ubel P, Stewart A. Sexual Harassment and Discrimination Experiences of Academic Medical Faculty. *JAMA* 2016; **315**: 2120–1.

28 Nelson RG, Rutherford JN, Hinde K, Clancy KBH. Signaling Safety: Characterizing Fieldwork Experiences and Their Implications for Career Trajectories. *Am Anthropol* 2017; **119**: 710–22.

29 Race- and gender-based bias persists in US science. *Nature* 2018; **554**: 561.

30 Babcock L, Recalde MP, Vesterlund L, Weingart L. Gender Differences in Accepting and Receiving Requests for Tasks with Low Promotability. *Am Econ Rev* 2017; **107**: 714–47.

31 Guarino CM, Borden VMH. Faculty Service Loads and Gender: Are Women Taking Care of the Academic Family? *Res High Educ* 2017; **58**: 672–94.

32 El-Alayli A, Hansen-Brown AA, Ceynar M. Dancing Backwards in High Heels: Female Professors Experience More Work Demands and Special Favor Requests, Particularly from Academically Entitled Students. *Sex Roles* 2018; : 1–15.

33 Antecol H, Bedard K, Stearns J. Equal but Inequitable: Who Benefits from Gender-Neutral Tenure Clock Stopping Policies? IZA Discussion Papers, 2016 http://ftp.iza.org/dp9904.pdf.

34 Sege R, Nykiel-Bub L, Selk S. Sex Differences in Institutional Support for Junior Biomedical Researchers. *JAMA* 2015; **314**: 1175–7.

35 Lerback J, Hanson B. Journals invite too few women to referee. *Nature* 2017; **541**: 455–7.

36 Hengel E. Publishing while Female: Gender Differences in Peer Review Scrutiny. 2016; published online Oct 1. http://www.erinhengel.com/research/publishing_female.pdf.

37 Larivière V, Ni C, Gingras Y, Cronin B, Sugimoto CR. Bibliometrics: global gender disparities in science. *Nature* 2013; **504**: 211–3.

38 Caplar N, Tacchella S, Birrer S. Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy* 2017; **1**: 0141.

39 Ghiasi G, Larivière V, Sugimoto CR. On the Compliance of Women Engineers with a Gendered Scientific System. *PLoS One* 2015; **10**: e0145931.

40 Budden AE, Tregenza T, Aarssen LW, Koricheva J, Leimu R, Lortie CJ. Double-blind review favours increased representation of female authors. *Trends Ecol Evol* 2008; **23**: 4–6.

41 Roberts SG, Verhoef T. Double-blind reviewing at EvoLang 11 reveals gender bias. *Journal of Language Evolution* 2016; **1**: 163–7.

42 Nittrouer CL, Hebl MR, Ashburn-Nardo L, Trump-Steele RCE, Lane DM, Valian V. Gender disparities in colloquium speakers at top universities. *Proceedings of the National Academy of Sciences* 2017; published online Dec 18. DOI:10.1073/pnas.1708414115.

43 Boiko JR, Anderson AJM, Gordon RA. Representation of Women Among Academic Grand Rounds Speakers. *JAMA Intern Med* 2017; **177**: 722–4.

44 Buell D, Hemmelgarn B, Straus SE. Proportion of women presenters at medical grand rounds at major academic centres in Canada, a retrospective observational study (in press). *BMJ Open*.

45 Files JA, Mayer AP, Ko MG, *et al.* Speaker Introductions at Internal Medicine Grand Rounds: Forms of Address Reveal Gender Bias. *J Womens Health* 2017; **26**: 413–9.

46 Wallis CJ, Ravi B, Coburn N, Nam RK, Detsky AS, Satkunasivam R. Comparison of postoperative outcomes among patients treated by male and female surgeons: a population based matched cohort study. *BMJ* 2017; **359**: j4366.

47 Sarsons H. Interpreting Signals in the Labor Market: Evidence from Medical Referrals. 2017. https://scholar.harvard.edu/files/sarsons/files/sarsons_jmp.pdf.

48 Jena AB, Khullar D, Ho O, Olenski AR, Blumenthal DM. Sex Differences in Academic Rank in US Medical Schools in 2014. *JAMA* 2015; **314**: 1149–58.

49 Carr PL, Raj A, Kaplan SE, Terrin N, Breeze JL, Freund KM. Gender Differences in Academic Medicine: Retention, Rank, and Leadership Comparisons From the National Faculty Survey. *Acad Med* 2018; published online Jan 30.

DOI:10.1097/ACM.0000000000002146.

50 Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J. Science faculty's subtle gender biases favor male students. *Proc Natl Acad Sci U S A* 2012; **109**: 16474–9.

51 MacNell L, Driscoll A, Hunt AN. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innov High Educ* 2015; **40**: 291–303.

52 Massen JJM, Bauer L, Spurny B, Bugnyar T, Kret ME. Sharing of science is most likely among male scientists. *Sci Rep* 2017; **7**: 12927.

53 Lutter M, Schröder M. Who becomes a tenured professor, and why? Panel data evidence from German sociology, 1980–2013. *Res Policy* 2016; **45**: 999–1013.

54 Breda T, Hillion M. Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France. *Science* 2016; **353**: 474–8.

55 Williams WM, Ceci SJ. National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. *Proc Natl Acad Sci U S A* 2015; **112**: 5360–5.

56 Ceci SJ, Williams WM. Women have substantial advantage in STEM faculty hiring, except when competing against more-accomplished men. *Front Psychol* 2015; **6**: 1532.

57 Leslie LM, Manchester CF, Dahm PC. Why and When Does the Gender Gap Reverse? Diversity Goals and the Pay Premium for High Potential Women. *Acad Manage J* 2017; **60**: 402–32.

58 Bohren JA, Imas A, Rosenberg M. The Dynamics of Discrimination: Theory and Evidence. 2017; published online Nov 18. DOI:10.2139/ssrn.3081873.

59 Bornmann L, Mutz R, Daniel H-D. Gender differences in grant peer review: A meta-analysis. *J Informetr* 2007; **1**: 226–38.

60 Tamblyn R, McMahon M, Girard N, Drake E, Nadigel J, Gaudreau K. Health services and policy research in the first decade at the Canadian Institutes of Health Research. *CMAJ Open* 2016; **4**: E213–21.

61 Ley TJ, Hamilton BH. Sociology. The gender gap in NIH grant applications. *Science* 2008; **322**: 1472–4.

62 Pohlhaus JR, Jiang H, Wagner RM, Schaffer WT, Pinn VW. Sex differences in application, success, and funding rates for NIH extramural programs. *Acad Med* 2011; **86**: 759–67.

63 Waisbren SE, Bowles H, Hasan T, *et al.* Gender differences in research grant applications and funding outcomes for medical school faculty. *J Womens Health* 2008; **17**: 207–14.

64 Bedi G, Van Dam NT, Munafo M. Gender inequality in awarded research grants. *Lancet* 2012; **380**: 474.

65 Head MG, Fitchett JR, Cooke MK, Wurie FB, Atun R. Differences in research funding for women scientists: a systematic comparison of UK investments in global infectious disease research during 1997-2010. *BMJ Open* 2013; **3**: e003362.

66 Svider PF, D'Aguillo CM, White PE, *et al.* Gender differences in successful National Institutes of Health funding in ophthalmology. *J Surg Educ* 2014; **71**: 680–8.

67 Mutz R, Bornmann L, Daniel H-D. Does Gender Matter in Grant Peer Review? *Zeitschrift für Psychologie* 2012; **220**: 121–9.

68 Albers CJ. Dutch research funding, gender bias, and Simpson's paradox. *Proceedings of the National Academy of Sciences* 2015; **112**: E6828–9.

69 Volker B, Steenbeek W. No evidence that gender contributes to personal research funding success in The Netherlands: A reaction to van der Lee and Ellemers. *Proc Natl Acad Sci U S A* 2015; **112**: E7036–7.

70 Ceci SJ, Williams WM. Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences* 2011; **108**: 3157–62.

71 Gino F, Wilmuth CA, Brooks AW. Compared to men, women view professional advancement as equally attainable, but less desirable. *Proc Natl Acad Sci U S A* 2015; **112**: 12354–9.

72 Canadian Institutes of Health Research. Foundation Grant – Stage 1 Application Instructions 4.1 Adjudication Criteria. 2017; published online April 5. http://www.cihr-irsc.gc.ca/e/48314.html#t2_4 (accessed Feb 21, 2018).

73 Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM. Mixed Effects Models and Extensions in Ecology with R. Springer, 2009.

74 Hubbard AE, Ahern J, Fleischer NL, *et al.* To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* 2010; **21**: 467–74.

75 van der Lee R, Ellemers N. Gender contributes to personal research funding success in The Netherlands. *Proc Natl Acad Sci U S A* 2015; **112**: 12349–53.

76 Malmström M, Johansson J, Wincent J. Gender Stereotypes and Venture Support Decisions: How Governmental Venture Capitalists Socially Construct Entrepreneurs' Potential. *Entrepreneurship Theory and Practice* 2017; published online Feb 1. DOI:10.1111/etap.12275.

77 Hyde JS, Mertz JE. Gender, culture, and mathematics performance. *Proc Natl Acad Sci U S A* 2009; **106**: 8801–7.

78 Goldin C, Rouse C. Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians. *Am Econ Rev* 2000; **90**: 715–41.

79 Steinpreis RE, Anders KA, Ritzke D. The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study. *Sex Roles* 1999; **41**: 509–28.

80 Blumenthal DM, Olenski AR, Yeh RW, *et al.* Sex Differences in Faculty Rank Among Academic Cardiologists in the United States. *Circulation* 2017; **135**: 506–17.

81 Okimoto TG, Brescoll VL. The price of power: power seeking and backlash against female

politicians. *Pers Soc Psychol Bull* 2010; **36**: 923–36.

82 Eagly AH, Karau SJ. Role congruity theory of prejudice toward female leaders. *Psychol Rev* 2002; **109**: 573–98.

83 Rudman LA, Moss-Racusin CA, Phelan JE, Nauts S. Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *J Exp Soc Psychol* 2012/1; **48**: 165–79.

84 McClean E, Martin SR, Emich K, Woodruff T. The social consequences of voice: An examination of voice type and gender on status and subsequent leader emergence. *Acad Manage J* 2017; published online Sept 14. DOI:10.5465/amj.2016.0148.

85 Zepeda L. The harassment tax. *Science* 2018; **359**: 126.

86 Pezzoni M, Mairesse J, Stephan P, Lane J. Gender and the Publication Output of Graduate Students: A Case Study. *PLoS One* 2016; **11**: e0145146.

87 Aakhus E, Mitra N, Lautenbach E, Joffe S. Gender and Byline Placement of Co-first Authors in Clinical and Basic Science Journals With High Impact Factors. *JAMA* 2018; **319**: 610–1.

88 West JD, Jacquet J, King MM, Correll SJ, Bergstrom CT. The role of gender in scholarly authorship. *PLoS One* 2013; **8**: e66212.

89 Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, Cassidy R. Sugimoto. Global Gender Disparities in Science. *Nature* 2013; **504**: 211–3.

90 StatisticsCanada. Number and salaries of full-time teaching staff at Canadian universities, 2016/2017. The Daily. 2017; published online April 25. http://www.statcan.gc.ca/daily-quotidien/170425/dq170425b-eng.pdf.

91 CIHR's Gender Equity Framework. 2017; published online Feb 10. http://www.cihr-irsc.gc.ca/e/50238.html (accessed Aug 2, 2017).

92 Tricco AC, Thomas SM, Antony J, *et al.* Strategies to Prevent or Reduce Gender Bias in Peer Review of Research Grants: A Rapid Scoping Review. *PLoS One* 2017; **12**: e0169718.

93 Unconscious Bias in Peer Review. http://www.cihr-irsc.gc.ca/lms/e/bias/ (accessed Aug 2, 2017).

94 Moss-Racusin CA, van der Toorn J, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J. A 'Scientific Diversity' Intervention to Reduce Gender Bias in a Sample of Life Scientists. *CBE Life Sci Educ* 2016; **15**. DOI:10.1187/cbe.15-09-0187.

95 Lai CK, Hoffman KM, Nosek BA. Reducing Implicit Prejudice. *Soc Personal Psychol Compass* 2013; **7**: 315–30.

96 Girod S, Fassiotto M, Grewal D, *et al.* Reducing Implicit Gender Leadership Bias in Academic Medicine With an Educational Intervention. *Acad Med* 2016; **91**: 1143–50.

97 Balafoutas L, Sutter M. Affirmative action policies promote women and do not harm

efficiency in the laboratory. *Science* 2012; **335**: 579–82.

98 Besley T, Folke O, Persson T, Rickne J. Gender Quotas and the Crisis of the Mediocre Man: Theory and Evidence from Sweden∗ . http://perseus.iies.su.se/~tpers/papers/Draft170122AER.pdf.

99 Hyll W. Gender Quotas and Human Capital Formation: A Relative Deprivation Approach. *Ger Econ Rev* 2017; **18**: 302–26.

100 Canadian Institutes of Health Research. 2017-2018 Foundation Grant competition: Summary of stage 1 results - CIHR. 2017; published online Dec 8. http://www.cihr-irsc.gc.ca/e/50740.html (accessed Feb 28, 2018).

101 Meerwijk EL, Sevelius JM. Transgender Population Size in the United States: a Meta-Regression of Population-Based Probability Samples. *Am J Public Health* 2017; **107**: e1–8.

102 Salzberg AJ. Removable Selection Bias in Quasi-Experiments. *Am Stat* 1999; **53**: 103–7.

APPENDIX 1. Methodological and Analytical Details

Population, Setting and Data

Beginning in 2014, the Canadian Institutes of Health Research (CIHR) phased out traditional open grant programs and divided all investigator-initiated funding into two new programs: the Project grant program and Foundation grant program. Both new programs used a staged review process in which lower-ranked applications were rejected from progressing to further stages. As in traditional programs, reviewers in the new Project grant program were instructed to primarily assess the research proposed. Seventy-five percent of the application score was based on reviewers' assessments of ideas and methods while 25% was based on reviewers' assessments of principal investigators' and teams' expertise, experience, and resources. In contrast, the Foundation grant program was about 'people, not projects' and was designed to provide grants to fund programs of research. At the first stage of the Foundation review process, reviewers were instructed to primarily assess the principal investigator, with 75% of the score allocated to reviewers' assessments of principal investigators' leadership, productivity, and the significance of their contributions, and 25% to a one-page summary of their proposed 5- or 7-year research program. Only principal investigators who passed this stage were invited to submit a detailed proposal describing their research. Thus, these new programs enabled a direct, quasi-experimental comparison of success rates of male and female applicants in grant programs with and without an explicit focus on the caliber of the principal investigator.

New investigators and those who had never held CIHR funding could apply to programs of their choice. Established principal investigators who already held CIHR funding were eligible for the Foundation program if one or more of their active CIHR grants was scheduled to end within a specific date range. Principal investigators could apply to multiple programs, with some restrictions. In the first cycle of the Foundation program, principal investigators who passed the first stage and were accepted to submit a full description of their research could not simultaneously apply to the last cycle of traditional programs. In the second cycle of the Foundation program, principal investigators could apply to Foundation and Project programs, providing they did not submit the same research proposal to both programs.

We analyzed data from all applications submitted to CIHR grant programs across all investigator-initiated competitions in 2011 through 2016. We excluded applications that were withdrawn, as these did not receive full peer review. We also excluded applications if the

principal investigator, referred to as the nominated principal applicant in the CIHR system, had not reported their sex, their age, the domain of research of their application, or if their self-reported age was unrealistic. We defined unrealistic ages as occurring when a principal investigator's self-reported birth year was prior to 1920 or after 2000. Ensuring correct date entry in web-based forms is a known challenge in human-computer interaction.[1] In the online system used to collect personal data in this study, the Canadian Common CV, the default birth date is prior to 1920, suggesting that such self-reported birth years were most likely to occur when people did not enter a birth date. Birth years in the 2000s were deemed to be errors and may have occurred, for example, when people accidentally entered the current year as they created or updated their Canadian Common CV rather than their birth year.

Data Source

Our dataset was a full export of CIHR competition data from its Electronic Information System (EIS). This dataset does not include withdrawn applications. It only includes applications submitted that were fully assessed by peer review and either approved or not.

Analysis

The binary outcome of interest was application success. In the model, we coded this as true (1) if the application was approved after the peer review process, and false (0) if not. Because our aim was to analyze the effects of peer review, we coded applications that were deemed fundable but not approved in the competition to which they were applied as unsuccessful, even if they were later awarded money through other administrative processes such as bridge grants or priority announcements for specific funding areas.

We modeled grant success rates as a function of the grant program, principal investigators' self-reported binary sex (male or female), self-reported age, self-declared domain of research (biomedical; clinical; health systems and services; or social, cultural, environmental and population health),[2] and an interaction term between each principal investigator's sex and the grant program to which they were applying. Grant programs were grouped into three categories: traditional investigator-initiated grant programs programs (includes regular open operating grant programs from fiscal years 2011/12 to 2014/15 which account for 88% of grants in the Traditional programs group plus 6 smaller programs, some of which continued into fiscal year 2015/16); Project (spring 2015 competition, fiscal year 2015/16); and Foundation (two competitions: 2014/15 and 2015/16). All predictors were categorical variables, except for applicant age, which was continuous and was mean-centered prior to analysis.

The interaction term allowed us to address the objective of this study by determining whether there was any effect of different review criteria on relative success rates between male and female applicants after controlling for age and domain of research. We controlled for these because younger cohorts of investigators included larger proportions of female principal investigators, as did domains of health research other than biomedical. The adjustment for age also helped account for the fact that both the Foundation program and Project program had a predefined minimal allocation to new investigators, meaning those within their first five years as independent investigators. The CIHR collected data about binary sex, not gender; therefore, our study assumes that people who self-reported as female or male identified as women or men, respectively. The CIHR did not collect complete data on other applicant characteristics that have been shown to be associated with disparities in funding and career progression; for example, career stage, race, ethnicity, Indigeneity and disability.[3,4]

We used Generalized Estimating Equations (GEE) to fit a logistic model that accounted for the same principal investigator submitting multiple applications,[5,6] including principal investigators who applied to both Project and Foundation programs. We conducted analyses in R statistical computing software, version 3.4.0,[7] using the geepack package to fit models.[8] We then used the fitted model to test the pairwise effect of sex within each program, using the lsmeans package.[9] This allowed us to compute marginal effects for specific contrasts of interest.

The R code used for our analyses is appended, with variable names masked for security and privacy reasons.

Results

There were a total of 25,706 applications during the five years of this study. We excluded 1,788 applications consisting primarily of principal investigators with unrealistic years of birth; i.e., birth years prior to 1920 (n=1,631) or after 2000 (n=12). The final dataset analyzed contained 23,918 applications from 7,093 unique principal investigators. There were 15,775 applications from 4,472 male principal investigators and 8,143 applications from 2,621 female principal investigators. Twenty-eight percent of principal investigators submitted a single application during the five-year study period, 20% submitted two applications, 25% submitted three or four applications, and the remaining 27% of principal investigators submitted five or more

applications. The maximum number of applications from a principal investigator during the 5-year period was 40.

The overall grant success rate across the data set was 15.8%. After adjusting for age and research domain, the predicted probability of funding was 0.9 percentage points higher for male principal investigators than female principal investigators in traditional programs (OR 0.934, 95% CI 0.854-1.022). This gap was 0.9 percentage points in the Project program (OR 0.998, 95% CI 0.794-1.229) and 4.0 percentage points in the Foundation program (OR 0.705, 95% CI 0.519-0.960). In other words, female applicants experienced significantly lower success rates than male applicants in Foundation, but not in Project nor in traditional programs. This was confirmed by using the model coefficients to compute contrasts and associated odds ratios for sexes within each program. Across all grant programs, odds of receiving funding were also lower in the three non-biomedical research domains and for younger principal investigators.

Tables S1 and Table S2 provide different ways of viewing these results. Table S1 shows the raw GEE results while Table S2 shows the observed success rates, predicted probabilities, and calculated odds ratios for the interaction between applicant sex and program at uniform values of age, averaged across domains.[9] Table S3 shows results by review stage for the cycles included in the quasi-experiment and thus in the analyses in the paper. Table S4 shows results by review stage for the following cycle after changes were made to the program, specifically, after a reviewer learning module on the topic of unconscious bias was implemented in the program.

TABLES

Table S1. Odds of Grant Success: GEE Results

	Odds Ratio (OR)	OR lower bound of 95% confidence interval	OR upper bound of 95% confidence interval
(Intercept): Male applicants in Program: traditional programs, Research Domain: Biomedical, Mean age	0.229	0.216	0.242
Female applicants	0.934	0.854	1.022
Program: Project	0.762	0.675	0.860
Program: Foundation	0.748	0.641	0.873
Age (mean-centered)	1.005	1.001	1.010
Research Domain: Clinical	0.815	0.738	0.900
Research Domain: Health Systems and Services	0.846	0.747	0.959
Research Domain: Population and Public Health	0.772	0.681	0.877
Sex * Program: Female applicants in Project	0.988	0.794	1.229
Sex * Program: Female applicants in Foundation	0.705	0.518	0.960

Table S2. Associations between Predictors and Funding Success

	Number of applications in data set	Percent successful (unadjusted)	Predicted probabilities* (after adjusting for age and research domain)	Odds Ratio (OR)*	OR lower bound of 95% confidence interval*	OR upper bound of 95% confidence interval*
Applicant Sex x Program Interaction						
Male applicants in traditional programs (reference)	11879	17.3%	16.3%	1.000		
Female applicants in traditional programs	6326	15.7%	15.4%	0.934	0.854	1.022
Male applicants in Project	2469	13.5%	12.9%	0.762	0.675	0.860
Female applicants in Project	1119	12.0%	12.1%	0.703	0.587	0.841
Male applicants in Foundation	1427	13.9%	12.7%	0.748	0.641	0.873
Female applicants in Foundation	698	9.2%	8.8%	0.493	0.375	0.647
Research Domain						
Biomedical (reference)	14159	16.9%	n/a	1.000		
Clinical	4497	14.3%	n/a	0.815	0.738	0.900
Health Systems and Services	2609	14.9%	n/a	0.846	0.747	0.959

Population and Public Health	2653	13.6%	n/a	0.772	0.681	0.877
Age						
Age (mean-centered)	23918	15.8%	n/a	1.005	1.001	1.010

*Predicted probabilities and odds ratios (including 95% confidence intervals) calculated with lsmeans package.

Table S3. Foundation results by review stage during study

Cycle	Stage	Male applicants: n (% of previous stages)	Female applicants: n (% of previous stages)
2014/15	Applied at Stage 1	850	490
Review focus: caliber of applicant	Invited to Stage 2 and applied	317 (317/850 = 37%)	126 (126/490 = 26%)
Review focus: quality of research	Funded	109 (109/317 = 34%, 109/850 = 13%)	41 (41/126 = 33%, 41/490 = 8%)
2015/16	Applied at Stage 1	623	279
Review focus: caliber of applicant	Invited to Stage 2 and applied	198 (198/623 = 32%)	62 (62/279 = 22%)
Review focus: quality of research	Funded	91 (91/198 = 46%, 91/623 = 15%)	29 (29/62 = 47%, 29/279 = 10%)

A subset of applications also underwent additional Stage 3 review after Stage 2, again with a focus on the quality of the research, before being funded or rejected. There were no differences by applicant's self-reported sex.

Table S4. Foundation results by review stage after implementation of unconscious bias reviewer training module

Cycle	Stage	Male applicants: n (% of previous stages)	Female applicants: n (% of previous stages)
2016/17	Applied at Stage 1	428	172
Review focus: caliber of applicant	Invited to Stage 2 and applied	165 (165/428 = 39%)	69 (69/172 = 40%)
Review focus: quality of research	Funded	54 (54/165 = 33%, 54/428 = 13%)	22 (22/69 = 32%, 22/172 = 13%)

REFERENCES

1 Bargas-Avila JA, Brenzikofer O, Tuch AN, Roth SP, Opwis K. Working towards Usable Forms on the World Wide Web: Optimizing Date Entry Input Fields. *Advances in Human-Computer Interaction* 2011; **2011**. DOI:10.1155/2011/202701.

2 The Four Themes of CIHR Funded Health Research - CIHR. 2014; published online Nov 24. http://www.cihr-irsc.gc.ca/e/48801.html (accessed July 8, 2017).

3 Ginther DK, Schaffer WT, Schnell J, *et al.* Race, ethnicity, and NIH research awards. *Science* 2011; **333**: 1015–9.

4 Hewitt T. Open Letter to University Presidents from the Canada Research Chairs Program. 2016; published online April 28. http://www.chairs-chaires.gc.ca/whats_new-quoi_de_neuf/2016/letter-lettre-eng.aspx (accessed July 7, 2017).

5 Zuur AF, Ieno EN, Walker N, Saveliev AA, Smith GM. Mixed Effects Models and Extensions in Ecology with R. Springer, 2009.

6 Hubbard AE, Ahern J, Fleischer NL, *et al.* To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology* 2010; **21**: 467–74.

7 R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2017 https://www.R-project.org/.

8 Halekoh U, Højsgaard S, Yan J, Others. The R package geepack for generalized estimating equations. *J Stat Softw* 2006; **15**: 1–11.

9 Lenth RV, Others. Least-squares means: the R package lsmeans. *J Stat Softw* 2016; **69**: 1–33.

APPENDIX 2. R Code

```
## INITIALIZE
rm(list=ls())                          # clear memory

setwd("M:/My Documents/_Documents/Gender Equity Analysis/3. Outputs")

## Data manipulation
library("tidyr")
library("dplyr")
library("lazyeval")
library("stringr")

## Data analysis & output
library("stargazer")

## Graphics
library("ggplot2")
library("GGally")
library("scales")

################################################################################
## LOAD DATA

## Read in competition data
CompData <- read.csv("../1. Raw data/export_competition_May_31_2017.csv")
##Data set includes information on applicants, applications, and funding status (approved grant amounts and durations).

##_____________________________________________________________________________
## Filters
## NOTE: withdrawn applications do not appear in this dataset.

## Keep only Investor-Initiated programs, with Program Codes: "MOP", "MCT", "FDN", "PJT", "PHE", "KRS", "KAL", "PP2", "PPP", "IPR";
##  remove 200711MOP and 201310MOP
open_grants <- c("MOP", "MCT", "FDN", "PJT", "PHE", "KRS", "KAL", "PP2", "PPP", "IPR")
OpenGrantData <- subset(CompData, program.code %in% open_grants)
## Exclude bridge MOP grants: basically, any MOP competition not ending in '....03MOP' or '....09MOP'
OpenGrantData <- subset(OpenGrantData, competition.code != "200711MOP" & competition.code != "201310MOP")
dim(OpenGrantData)

## Keep only Fiscal Years since 2011/06 (to 201516); 2016 foundation competition was pending at the time of analysis.
sort(unique(OpenGrantData$YEAR.fiscal))
OpenGrantData <- subset(OpenGrantData, YEAR.fiscal >= 201112 & YEAR.fiscal < 201600)
```

```
dim(OpenGrantData)

## Remove unknown levels of factors
OpenGrantData <- filter(OpenGrantData, SEX.npi != "N", !THEME.code %in% c("N", ""))    #, INSTITUTE.code != "I999")

##_____________________________________________________________________________
## Calculate extra columns

  OpenGrantData <- mutate(OpenGrantData, DOB.npi = as.Date(DOB.npi, format = "%d-%b-%Y"))
  dob.years <- as.numeric( format(OpenGrantData$DOB.npi, "%Y") )
  age.clean <- OpenGrantData$AGE.npi
  table(dob.years)
  table(age.clean)
  age.clean[dob.years < 1920] <- NA
  age.clean[dob.years > 2000] <- NA
  OpenGrantData <- mutate(OpenGrantData, AGE.npi = age.clean)

OpenGrantData <- mutate(OpenGrantData, AGE.npi = ifelse(AGE.npi >= 100, NA, AGE.npi))

## Grant Duration (total approved grant length)
OpenGrantData <- mutate(OpenGrantData, DURATION.approved = (approved.years * 12) + approved.months) # in MONTHS

leadership_programs <- c("FDN")
OpenGrantData <-
  mutate(OpenGrantData, leadership.comp = program.code %in% leadership_programs)

lapply(open_grants, function (prog) table(filter(CompData, program.code == prog)$competition.code %>% as.character()))
lapply(open_grants, function (prog) table(filter(OpenGrantData, program.code == prog)$competition.code %>% as.character()))
trans.OOGP <- c("201503MOP")
prog_bins <- c(ref = "OOGP + other", trans.oogp = "Trans. OOGP", PJT = "PJT", FDN = "FDN") # Older
prog_bins <- prog_bins[names(prog_bins) != 'trans.oogp']
prog_bins_labels <- sub(" ", "\n", prog_bins) # for plotting labels
OpenGrantData <-
  mutate(OpenGrantData,
         leadership.comp = program.code %in% leadership_programs,
         ##program.bin = ifelse(competition.code %in% trans.OOGP, prog_bins['trans.oogp'],
         program.bin = ifelse(program.code %in% c("FDN", "PJT"), as.character(program.code), prog_bins['ref']),
         program.bin = factor(program.bin, levels = prog_bins) # force order
         )

OpenGrantData <- subset(OpenGrantData, !is.na(AGE.npi))
OpenGrantData <- droplevels(OpenGrantData)
dim(OpenGrantData)
```

```
##_____________________________________________________________________________
## Approved Grants for analysis of duration, amounts, etc.
ApprovedGrants <- OpenGrantData %>% filter(application.approved, DURATION.approved > 0, AMOUNT.approved > 0)

OpenGrantData %>% filter(application.approved) %>% with( table(DURATION.approved) ) # 4 0s
OpenGrantData %>% filter(application.approved) %>% with( length(which(AMOUNT.approved == 0)) )
OpenGrantData %>% filter(application.approved) %>% with( length(which(AMOUNT.requested == 0)))

nrow(OpenGrantData %>% filter(application.approved)) - nrow(ApprovedGrants)

## Graphs of main effects of interest

legend.name.sex <- "Sex of Investigator"
scale.sex.colour <-
  scale_colour_manual(name = legend.name.sex,
                      values = c(F = "#330066", M = "#66CCFF"),
                      breaks = c("F", "M"),
                      labels = c("Female", "Male")
                      )

scale.sex.fill <-
  scale_fill_manual(name = legend.name.sex,
                    values = c(F = "#330066", M = "#66CCFF"),
                    breaks = c("F", "M"),
                    labels = c("Female", "Male")
                    )

scale.sex.size <-
  scale_size_manual(name = legend.name.sex,
                    values = c(F = 1.5, M = 1),
                    breaks = c("F", "M"),
                    labels = c("Female", "Male")
                    )

scale.leadership.lty <-
  scale_linetype_manual(name = "Leadership \ncomponent \nof peer review",
                     values = c('TRUE' = "42", 'FALSE' = "solid"),
                     breaks = c(TRUE, FALSE),
                     labels = c("Yes", "No")
                     )


## Prepare summary data for graphs
```

```
summarize_grant_data <- function (.data = OpenGrantData, ...)
{
  .groups <- lazyeval::lazy_dots(...)
  .summary <- group_by_(.data, .dots = .groups) %>%
    summarize(n = n(),
              n.approved = length(which(application.approved)),
              success.rate = n.approved / n,
              age.avg = mean(AGE.npi),
              age.sd = sd(AGE.npi)
              )
  .summary <-
    full_join(.summary,
            filter(.data, application.approved, DURATION.approved > 0, AMOUNT.approved > 0) %>%
            group_by_(.dots = .groups) %>%
            summarize(duration.avg = mean(DURATION.approved),
                      amount.avg = mean(AMOUNT.approved),
                      amount.sd  = sd(AMOUNT.approved),
                      amount.per.yr = mean(AMOUNT.approved / DURATION.approved),
                      amount.per.yr.sd = sd(AMOUNT.approved / DURATION.approved),
                      requested.avg = mean(AMOUNT.requested),
                      req.decided.avg = mean(AMOUNT.approved / AMOUNT.requested)
                      )
            )
  .summary
}


##==============================================================
## Overall Summaries

dim(OpenGrantData)
length(unique(OpenGrantData$application.id)) # number of applications (should be number of rows)
length(unique(OpenGrantData$npi.id)) # number of unique applicants

filter(OpenGrantData, SEX.npi == "M")$application.id %>% unique() %>% length()
filter(OpenGrantData, SEX.npi == "M")$npi.id %>% unique() %>% length()
filter(OpenGrantData, SEX.npi == "F")$application.id %>% unique() %>% length()
filter(OpenGrantData, SEX.npi == "F")$npi.id %>% unique() %>% length()

sort(table(OpenGrantData$npi.id))

summarize_grant_data(OpenGrantData, SEX.npi)

Data.summary <- summarize_grant_data(OpenGrantData, YEAR.fiscal, SEX.npi)
```

```
str(Data.summary)

npi.plot <- ggplot(as.data.frame(table(OpenGrantData$npi.id)), aes(x = Freq, y = ..density..)) +
  geom_histogram(binwidth = 1) +
  ##geom_density() +
  labs(title = "Number of applications per principal investigator",
       subtitle = "Fiscal years 2011/12 to 2015/16", x = "Number of applications", y = "proportion (density)") +
  theme_minimal()
npi.plot

npi.cumu.plot <- ggplot(as.data.frame(sort(table(OpenGrantData$npi.id))), aes(x = Freq)) +
  stat_ecdf() +
  labs(title = "Number of applications per principal investigator",
       subtitle = "Fiscal years 2011/12 to 2015/16", x = "Number of applications", y = "Cumulative proportion") +
  theme_minimal()
npi.cumu.plot

age.plot <- ggplot(OpenGrantData, aes(x = AGE.npi)) +
  geom_histogram(aes(y = ..density..), binwidth = 1) +
  stat_ecdf() +
  labs(title = "Age of applicant (at competition date)",
       subtitle = "Fiscal years 2011/12 to 2015/16", x = "Age", y = "proportion (density)") +
  theme_minimal()
age.plot



##==============================================================
## Summaries with Leadership component (FDN)

Summary.l <- summarize_grant_data(OpenGrantData, YEAR.fiscal, leadership.comp, SEX.npi)
str(Summary.l)

## DV1 = application.approved

l.success.yr.plot <- ggplot(Summary.l,
                       aes(x = as.integer(YEAR.fiscal / 100), y = success.rate *100)) +
geom_line(aes(colour = SEX.npi, lty = leadership.comp)) +
scale.sex.colour + scale.sex.size +
scale.leadership.lty +
labs(x = "Fiscal Year", y = "Success rate (%)") +
theme_minimal()
l.success.yr.plot
```

```
l.success.plot <- ggplot(summarize_grant_data(OpenGrantData, leadership.comp, SEX.npi),
                        aes(x = leadership.comp, y = success.rate *100, fill = SEX.npi) ) +
geom_bar(stat = "identity", position = position_dodge()) +
scale.sex.fill +
labs(title = "Success rate (FY 2011-2015)", x = "Leadership component", y = "Success rate (%)") +
theme_minimal()
l.success.plot


## application.approved by sex, theme, leader, fiscal year
Summary.pillar.l <- summarize_grant_data(OpenGrantData, THEME.en, YEAR.fiscal, leadership.comp, SEX.npi)

l.success.yr.pillar.plot <- l.success.yr.plot %+% Summary.pillar.l +
  facet_wrap(~ THEME.en)
print(l.success.yr.pillar.plot)

l.success.pillar.plot <- l.success.plot %+% summarize_grant_data(OpenGrantData, THEME.en, leadership.comp, SEX.npi) +
  facet_wrap(~ THEME.en)
print(l.success.pillar.plot)


## application.approved by sex, AGE.group, leadership, fiscal year

Summary.age.l <-
  mutate(OpenGrantData, AGE.group = cut(AGE.npi, breaks = c(0, 40, 50, 60, 100), right = FALSE)) %>%
  summarize_grant_data(AGE.group, YEAR.fiscal, leadership.comp, SEX.npi)

l.success.yr.age.plot <- l.success.yr.plot %+% Summary.age.l +
  facet_wrap(~ AGE.group) +
  labs(caption = "Age bins (intervals): [40, 50) indicates '40 to 49'")
print(l.success.yr.age.plot)

l.success.age.plot <- l.success.plot %+%
  (mutate(OpenGrantData, AGE.group = cut(AGE.npi, breaks = c(0, 40, 50, 60, 100), right = FALSE)) %>%
  summarize_grant_data(AGE.group, leadership.comp, SEX.npi)) +
  facet_wrap(~ AGE.group) +
  labs(caption = "Age bins (intervals): [40, 50) indicates '40 to 49'")
print(l.success.age.plot)


## DV2 = AMOUNT.approved
## Include DURATION.approved, and funding requested

## Grant Amount
```

```
amt.scale <- 100000
l.amount.plot <- ggplot(Summary.l,
                        aes(x = as.integer(YEAR.fiscal / 100), y = amount.avg / amt.scale,
                            ymin = (amount.avg - amount.sd) / amt.scale, ymax = (amount.avg + amount.sd) / amt.scale)
                        ) +
geom_ribbon(aes(fill = SEX.npi, lty = leadership.comp), alpha = 0.1) +
geom_line(aes(colour = SEX.npi, lty = leadership.comp)) +
scale.sex.colour + scale.sex.size + scale.sex.fill +
scale.leadership.lty +
labs(x = "Fiscal Year", y = "Grant amount (x $100,000)") +
theme_minimal()
print(l.amount.plot)

## Grant Duration
l.duration.plot <- l.amount.plot +
  aes(y = duration.avg / 12, ymin = NULL, ymax = NULL) +
  scale_y_continuous(limits = c(0, 10), breaks = seq(0, 10, by = 2)) +
  labs(y = "Grant duration (years)")
l.duration.plot$layers <- l.duration.plot$layers[2]
print(l.duration.plot)

## Grant Amount / Year
l.amount.per.yr.plot <- l.amount.plot +
  aes(y = amount.per.yr / amt.scale,
      ymin = (amount.per.yr - amount.per.yr.sd) / amt.scale,
      ymax = (amount.per.yr + amount.per.yr.sd) / amt.scale ) +
  labs(y = "Amount per year (x $100,000)")
print(l.amount.per.yr.plot)



##==============================================================
## Summaries by Program bins

Summary.b <- summarize_grant_data(OpenGrantData, program.bin, SEX.npi) #%>%
  ##ungroup() %>% mutate(program.bin = factor(program.bin, levels = prog_bins, labels = prog_bins_labels))
Summary.b.model <- summarize_grant_data(OpenGrantData, program.bin, SEX.npi, THEME.code)

relabel_prog_bins <- function (.data, levels = prog_bins, labels = prog_bins_labels)
{
  .data <- .data %>% ungroup()
  if ("program.bin" %in% colnames(.data))
    .data <- .data %>% mutate(program.bin = factor(program.bin, levels = levels, labels = labels))
  if ("SEX.npi" %in% colnames(.data))
```

```
    .data <- .data %>% mutate(SEX.npi = relevel(SEX.npi, ref = "M"))
  .data
}

## application.approved by sex
b.success.plot <- ggplot(Summary.b %>% relabel_prog_bins,
                        aes(x = program.bin, y = success.rate *100, fill = SEX.npi) ) +
geom_bar(stat = "identity", position = position_dodge()) +
scale.sex.fill +
labs(title = "Success rate by program group (FY 2011-2015)", x = "Program Group", y = "Success rate (%)") +
theme_minimal()
b.success.plot

## application.approved by sex, theme
Summary.pillar.b <- summarize_grant_data(OpenGrantData, THEME.en, program.bin, SEX.npi)

b.success.pillar.plot <- b.success.plot %+% (Summary.pillar.b %>% relabel_prog_bins()) +
  facet_wrap(~ THEME.en)
b.success.pillar.plot

## application.approved by sex, AGE.group
Summary.age.b <-
  mutate(OpenGrantData, AGE.group = cut(AGE.npi, breaks = c(0, 40, 50, 60, 100), right = FALSE)) %>%
  summarize_grant_data(AGE.group, program.bin, SEX.npi)

b.success.age.plot <- b.success.plot %+% (Summary.age.b %>% relabel_prog_bins()) +
  facet_wrap(~ AGE.group) +
  labs(caption = "Age bins (intervals): [40, 50) indicates '40 to 49'")
b.success.age.plot


## DV2 = AMOUNT.approved
## Include DURATION.approved, and funding requested

## Grant Amount
amt.scale <- 100000
b.amount.plot <- ggplot(Summary.b %>% relabel_prog_bins(),
                        aes(x = program.bin, y = amount.avg / amt.scale,
                            ymin = (amount.avg - amount.sd) / amt.scale, ymax = (amount.avg + amount.sd) / amt.scale)
                        ) +
geom_col(aes(fill = SEX.npi), position = "dodge") +
geom_errorbar(aes(group = SEX.npi), position = position_dodge(width = 0.9), width = 0.2, colour = "#666666") +
scale.sex.fill +
labs(x = NULL, y = "Grant amount (x $100,000)", caption = "error bars indicate 1 sd.") +
```

```
theme_minimal()
b.amount.plot

b.amount.boxplot <- ggplot(ApprovedGrants %>% relabel_prog_bins(),
                       aes(x = program.bin, y = AMOUNT.approved / amt.scale)
                       ) +
geom_boxplot(aes(colour = SEX.npi), position = "dodge") +
scale.sex.colour +
labs(x = NULL, y = "Grant amount (x $100,000)") +
theme_minimal()
b.amount.boxplot

b.amount.log.boxplot <- b.amount.boxplot + scale_y_log10()
b.amount.log.boxplot

## Grant Duration
b.duration.plot <- b.amount.plot +
  aes(y = duration.avg / 12, ymin = NULL, ymax = NULL) +
  scale_y_continuous(limits = c(0, 10), breaks = seq(0, 10, by = 2)) +
  labs(y = "Grant duration (years)")
b.duration.plot$layers <- b.duration.plot$layers[1]
b.duration.plot

b.duration.boxplot <- b.amount.boxplot +
  aes(y = DURATION.approved / 12) +
  scale_y_continuous(limits = c(0, 10), breaks = seq(0, 10, by = 2)) +
  labs(y = "Grant duration (years)")
b.duration.boxplot

## Grant Amount / Year
b.amount.per.yr.plot <- b.amount.plot +
  aes(y = amount.per.yr / amt.scale,
      ymin = (amount.per.yr - amount.per.yr.sd) / amt.scale,
      ymax = (amount.per.yr + amount.per.yr.sd) / amt.scale ) +
  labs(y = "Amount per year (x $100,000)")
b.amount.per.yr.plot

b.amount.per.yr.boxplot <- b.amount.boxplot  +
  aes(y = (AMOUNT.approved / DURATION.approved) / amt.scale) +
labs(y = "Amount per year (x $100,000)") +
theme_minimal()
b.amount.per.yr.boxplot

b.amount.per.yr.log.boxplot <- b.amount.per.yr.boxplot + scale_y_log10()
```

```
b.amount.per.yr.log.boxplot




################################################################################
## DATA ANALYSIS

## Variables
## Dependent variables:
##  DV1 = application.approved         funded/not funded (binary), primary outcome
##  DV2 = AMOUNT.approved   funding level (continuous, many 0s), secondary outcome
## Independent variables:
##  IV1 = SEX.npi               sex of the nominated principal investigator ('applicant') (male, female, unknown)
##  IV2 = leadership.comp       competition type (with / without leadership component); 2014 and later ONLY
## Covariates:
##  COV1 = AGE.npi              age of NPI (continuous or binned)
##  COV2 = THEME.code     (primary) pillar of application (categorical: B, C, H, S; N)
##  COV3 = INSTITUTE.code (primary) institute of application (categorical: 13 values)
##  DURATION.approved?

##==============================================================================
## Prepare data for analysis

## Center Age (the intercept is interpreted relative to the average Age, rather than an Age of 0)
OpenGrantData <- mutate(OpenGrantData, AGE_C = AGE.npi - mean(AGE.npi, na.rm = TRUE))

## Convert fiscal years to regular numbers (so differences are 1); first year == 0 (so intercept is relative to this year).
OpenGrantData <- mutate(OpenGrantData, YEAR_F = as.integer(YEAR.fiscal / 100) )
OpenGrantData <- mutate(OpenGrantData, YEAR_F = YEAR_F - min(YEAR_F) )

## reorder other factor levels to put reference levels first
OpenGrantData <- mutate(OpenGrantData,
                        SEX.npi = relevel(SEX.npi, ref = "M"), # relative to males
                        THEME.code = relevel(THEME.code, ref = "B") # relative to Biomedical
                        )

## Data must be sorted by Applicant for gee function; also sort by FY within Applicant (put applications in chronological order).
OpenGrantData <- mutate(OpenGrantData, Date.Competition = as.Date(Date.Competition, format = "%d-%b-%Y")) %>%
  arrange(npi.id, Date.Competition)
```

```
##==============================================================================
## Model-fitting: LEADERSHIP

library(geepack)
library(car)

OpenGrants.lm <- lm(application.approved ~ SEX.npi + leadership.comp + YEAR_F + AGE.npi + THEME.code, data=OpenGrantData)
vif(OpenGrants.lm)


## GEE model
## Try regular glm first: p-values will be too low
OpenGrants.l.glm <- glm(application.approved ~ SEX.npi * leadership.comp + YEAR_F + AGE_C + THEME.code, data=OpenGrantData, family =
binomial)
summary(OpenGrants.l.glm)
anova(OpenGrants.l.glm)
anova(OpenGrants.l.glm, update(OpenGrants.l.glm, . ~ . - SEX.npi:leadership.comp))
drop1(OpenGrants.l.glm, test = "Chi")  # All terms appear significant


## run bivariate analyses first, with only 1 IV in the model
GEE.sex <- geeglm(application.approved ~ SEX.npi, id = npi.id, data = OpenGrantData, family = binomial, corstr="exchangeable")
summary(GEE.sex)

GEE.lead <- update(GEE.sex, application.approved ~ leadership.comp)
summary(GEE.lead)

GEE.fy <- update(GEE.sex, application.approved ~ YEAR_F)
summary(GEE.fy)

GEE.age <-update(GEE.sex, application.approved ~ AGE.npi)
summary(GEE.age)

GEE.pillar <-update(GEE.sex, application.approved ~ THEME.code)
summary(GEE.pillar)



## Check if FY interacts with SEX.npi
GEE.fy.sex <- geeglm(application.approved ~ SEX.npi * YEAR_F, id = npi.id,
                     data = filter(OpenGrantData, YEAR_F > "200506" ), family = binomial, corstr = "exchangeable")
GEE.fy_sex <- update(GEE.fy.sex, application.approved ~ SEX.npi + YEAR_F) # update the model with no interaction term
summary(GEE.fy.sex)
anova(GEE.fy.sex)
```

```
anova(GEE.fy_sex)

if (T) {
  anova(GEE.fy.sex, GEE.fy_sex)           # dropping interaction: no significant difference
  anova(GEE.fy_sex, update(GEE.fy.sex, application.approved ~ SEX.npi))    # drop Year: no significant difference.
  anova(GEE.fy_sex, update(GEE.fy.sex, application.approved ~ YEAR_F))      # drop SEX.npi is significant
}




## Then run multivariate analysis with IV1 + IV2 + IV3 etc
## The working correlation matrix ("independence", "exchangeable", etc.) could be changed, for example to exchangeable in order to find
the one that suits best the data

GEE.full <-
  geeglm(application.approved ~ SEX.npi * leadership.comp + AGE_C + THEME.code + YEAR_F,
         id = npi.id, data = OpenGrantData, family = binomial, corstr="exchangeable")
summary(GEE.full)
anova(GEE.full)

GEE.nolead <- update(GEE.full, . ~ . - leadership.comp - SEX.npi:leadership.comp)
summary(GEE.nolead)
anova(GEE.nolead)



OpenGrants.gee <- GEE.full
summary(OpenGrants.gee)
anova(OpenGrants.gee)

##==============================================================================
## Model fitting: program bins
## Check for collinearity
OpenGrants.b.lm <- lm(application.approved ~ SEX.npi + program.bin + AGE_C + THEME.code + YEAR_F, data=OpenGrantData)
vif(OpenGrants.b.lm)                   # program.bin is somewhat collinear with YEAR.fiscal: PJT and FDN are not in all Years.

## Try regular glm first: p-values will be too low (risk of Type I error)
OpenGrants.b.glm <- glm(application.approved ~ SEX.npi * program.bin + AGE_C + THEME.code + YEAR_F, data = OpenGrantData, family =
binomial)
summary(OpenGrants.b.glm)
anova(OpenGrants.b.glm)
anova(OpenGrants.b.glm, update(OpenGrants.b.glm, . ~ . - SEX.npi:program.bin), test = "Chi")
drop1(OpenGrants.b.glm, test = "Chi")
```

```
## Univariate model
GEE.prog <- geeglm(application.approved ~ program.bin, id = npi.id, data = OpenGrantData, family = binomial, corstr="exchangeable")
summary(GEE.prog)


## GEE ================================
GEE.b.full <-
  geeglm(application.approved ~ SEX.npi * program.bin + AGE_C + THEME.code,
         id = npi.id, data = OpenGrantData, family = binomial, corstr = "exchangeable")
summary(GEE.b.full)
anova(GEE.b.full)
stats::coef(GEE.b.full)

GEE.b.full.yr <- update(GEE.b.full, . ~ . + YEAR_F)
summary(GEE.b.full.yr)
anova(GEE.b.full.yr)

GEE.b.ageint <- update(GEE.b.full, . ~ . + SEX.npi:AGE_C)
summary(GEE.b.ageint)
anova(GEE.b.ageint)

GEE.b.noint <- update(GEE.b.full, . ~ . - SEX.npi:program.bin)
summary(GEE.b.noint)
summary(update(GEE.b.full.yr, . ~ . - SEX.npi:program.bin))

GEE.b.int.psex   <- update(GEE.b.full, . ~ . + SEX.npi:THEME.code)
summary(GEE.b.int.psex)                # interaction NS, but SEX is marginally signif.; other coef. similar
anova(GEE.b.int.psex)                  # 2-way interaction NS
GEE.b.int.pprog  <- update(GEE.b.full, . ~ . + program.bin:THEME.code)
summary(GEE.b.int.pprog)               # Minimal effect on other coefficients
anova(GEE.b.int.pprog)                 # 2-way interaction NS
GEE.b.int.pillar <- update(GEE.b.full, . ~ . + SEX.npi * program.bin * THEME.code)
summary(GEE.b.int.pillar)              # SEX NS again; Effect of sex within FDN now NS!
anova(GEE.b.int.pillar)




##==============================================================================
## Model checking

num_clusters <- length(unique(OpenGrantData$npi.id))
message("Number of unique applicants (\'Number of clusers\' in GEE):    ", num_clusters)
```

```
max_cluster_size <- max( table(OpenGrantData$npi.id) )
message("Largest number of applications from a single NPI (\'Maximum cluster size\' in GEE):    ", max_cluster_size)

b.res <- residuals(GEE.b.full, type = "pearson")
plot(b.res)
qqnorm(b.res)
plot(predict(GEE.b.full), b.res)
probabilityAxis()
plot(GEE.b.full$y, b.res)
probabilityAxis()

## Predicted vs. observed
plot(GEE.b.full$y, predict(GEE.b.full))
probabilityAxis()

## See graphs below comparing observed vs. model-predicted success rates.



##==============================================================================
## Results
## back-transform logit values to odds: exp(logit)
## backtransform logit to probabilities: plogis(logit)

GEE.b.s <- summary(GEE.b.full)
GEE.b.coef <- coef(summary(GEE.b.full))
GEE.b.anova <- anova(GEE.b.full)

## Coefficients on odds scale (multiplicative effect on odds relative to intercept: i.e., Odds Ratio (OR)):
exp(coef(GEE.b.full))

## Coefficients and standard errors on Odds scale:
exp(GEE.b.coef[1:2])
GEE.b.cl <- GEE.b.coef[1:2]
CI.mult <- qnorm((1 + 0.95) / 2)       # 1.96
GEE.b.cl <- mutate(GEE.b.cl, LCL.95 = Estimate - (CI.mult * Std.err), UCL.95 = Estimate + (CI.mult * Std.err))
rownames(GEE.b.cl) <- rownames(GEE.b.coef)
GEE.b.OR.cl <- exp(GEE.b.cl[-2])
GEE.b.OR.cl <- rename(GEE.b.OR.cl, OR = Estimate)



##_____________________________________________________________________________
## Multiple Comparisons (sex differences between Programs)
```

```
library("estimability")
library(lsmeans)
(b.posthoc.sex.prog <- pmmeans(GEE.b.full, pairwise ~ SEX.npi * program.bin))
(b.posthoc.sex      <- pmmeans(GEE.b.full, pairwise ~ SEX.npi | program.bin))
(b.posthoc.prog     <- pmmeans(GEE.b.full, pairwise ~ program.bin | SEX.npi))

pmmeans(GEE.b.full.yr, pairwise ~ SEX.npi | program.bin)

pmmeans(GEE.b.int.pillar, pairwise ~ SEX.npi | program.bin)


## Simpler model of only the terms of interest



b.sex.prog.pmmeans <- transform(summary(b.posthoc.sex.prog$pmmeans),
                                SEX.npi = relevel(SEX.npi, ref = "M"), # to match data
                                p.mean = plogis(pmmean),
                                p.LCL = plogis(asymp.LCL),
                                p.UCL = plogis(asymp.UCL))

b.sex.prog.contrasts <- transform(summary(b.posthoc.sex$contrasts),
                                  OR = exp(estimate),
                                  OR.LCL = exp(estimate - (CI.mult * SE)),
                                  OR.UCL = exp(estimate + (CI.mult * SE)),
                                  OR.F = exp(-estimate),
                                  OR.F.LCL = exp(-estimate - (CI.mult * SE)),
                                  OR.F.UCL = exp(-estimate + (CI.mult * SE))
                                  )

b.sex.prog.int.contr <- transform(summary(b.posthoc.sex.prog$contrasts),  # full interaction results
                                  OR = exp(estimate),
                                  OR.LCL = exp(estimate - (CI.mult * SE)),
                                  OR.UCL = exp(estimate + (CI.mult * SE)),
                                  OR.F = exp(-estimate),
                                  OR.F.LCL = exp(-estimate - (CI.mult * SE)),
                                  OR.F.UCL = exp(-estimate + (CI.mult * SE))
                                  )




##_____________________________________________________________________________
## Publication graphs of observed & predicted (with CIs)
```

```
progs2pub <- function (.data, lvls = levels(.data$program.bin), labels = NULL)
{
  # Better labels for publication graphs
  if (is.null(labels)) {
    labels <- lvls
    labels <- gsub("OOGP.*", "Traditional\n OOGP", labels)
    labels <- gsub("^P.*", "Project", labels)
    labels <- gsub("^F.*", "Foundation", labels)
  }
  .data <- .data %>% ungroup()
  if ("program.bin" %in% colnames(.data))
    .data <- .data %>% mutate(program.bin = factor(program.bin, levels = lvls, labels = labels))
  if ("SEX.npi" %in% colnames(.data))
    .data <- .data %>% mutate(SEX.npi = relevel(SEX.npi, ref = "M"))
  .data
}


dodge.width <- 0.8

success.pred.sex.prog.plot <-
  ggplot(Summary.b %>% progs2pub(),
         aes(x = program.bin, y = success.rate *100, fill = SEX.npi)
         ) +
geom_col(position = position_dodge(width = dodge.width), width = dodge.width) +
geom_point(aes(y = p.mean * 100, group = SEX.npi), colour = "#666666",  # Predicted marginal means from the model
           data = b.sex.prog.pmmeans %>% progs2pub(),
           position = position_dodge(width = dodge.width), show.legend = FALSE) +
geom_errorbar(aes(y = p.mean * 100, ymin = p.LCL * 100, ymax = p.UCL * 100, group = SEX.npi),
              data = b.sex.prog.pmmeans %>% progs2pub(),
              position = position_dodge(width = dodge.width), width = 0.5,
              colour = "#666666") +
scale_y_continuous(limits = c(0, 20), expand = c(0, 0)) +
##scale.sex.fill +      # full version below
scale_fill_manual(name = "Sex of \nInvestigator",
                  values = c(F = "#330066", M = "#66CCFF"),
                  breaks = c("F", "M"),
                  labels = c("Female", "Male")
                  ) +
labs(title = "Success rate by program group", subtitle =  "FY 2011/12 to 2015/16",
     caption = "Points and error bars indicate model-predicted marginal means\n and 95% confidence intervals",
     x = NULL, y = "Success rate (%)") +
theme_classic()
```

```
success.pred.sex.prog.plot

progs2pub <- function (.data, lvls = levels(.data$program.bin), labels = NULL)
{
  if (is.null(labels)) {
    labels <- lvls
    labels <- gsub("OOGP.*", "Traditional", labels) #HW changed label
    #HW note that I am not 100% certain that this won't affect other code
    labels <- gsub("^P.*", "Project", labels)
    labels <- gsub("^F.*", "Foundation", labels)
  }
  .data <- .data %>% ungroup()
  if ("program.bin" %in% colnames(.data))
    .data <- .data %>% mutate(program.bin = factor(program.bin, levels = lvls, labels = labels))
  if ("SEX.npi" %in% colnames(.data))
    .data <- .data %>% mutate(SEX.npi = relevel(SEX.npi, ref = "M"))
  .data
}


dodge.width <- 0.8

success.pred.sex.prog.plot <-
  ggplot(Summary.b %>% progs2pub(),
         aes(x = program.bin, y = success.rate *100, fill = SEX.npi)
  ) +
  geom_col(position = position_dodge(width = dodge.width), width = dodge.width) +
  geom_point(aes(y = p.mean * 100, group = SEX.npi), colour = "#666666",  # Predicted marginal means from the model
             data = b.sex.prog.pmmeans %>% progs2pub(),
             position = position_dodge(width = dodge.width), show.legend = FALSE) +
  geom_errorbar(aes(y = p.mean * 100, ymin = p.LCL * 100, ymax = p.UCL * 100, group = SEX.npi),
                data = b.sex.prog.pmmeans %>% progs2pub(),
                position = position_dodge(width = dodge.width), width = 0.5,
                colour = "#666666") +
  scale_y_continuous(limits = c(0, 20), expand = c(0, 0)) +
  ##scale.sex.fill +      # full version below
  scale_fill_manual(name = "Sex of \nApplicant", #HW changed Investigator to Applicant
                    values = c(F = "#a1d99b", M = "#31a354"), #HW changed colours
                    breaks = c("F", "M"),
                    labels = c("Female", "Male")
  ) +
  labs(x = NULL, y = "Success rate (%)") + #HW removed title & caption
  theme_classic()
success.pred.sex.prog.plot
```

```
PmeansData <- b.sex.prog.pmmeans %>% progs2pub()
PmeansData$PointValues <- PmeansData$p.mean*100

tiff("hollyplot1.tif", units="in", width=6, height=6, res=300)
success.pred.sex.prog.plot
dev.off()


## Observed OR's
sex.prog.or <- Summary.b %>% select(program.bin, SEX.npi, success.rate) %>%
  spread(key = SEX.npi, val = success.rate) %>%
  mutate(odds.F = F / (1 - F), odds.M = M / (1 - M), OR = odds.F / odds.M, OR.fill = 1 - OR)

success.pred.sex.diff.plot <-
  ggplot(b.sex.prog.contrasts %>% progs2pub(), aes(x = program.bin, y = OR.F) ) +
  ##geom_col(aes(y = OR), data = sex.prog.or %>% relabel_prog_bins(),
  ##         colour = "#666666", width = dodge.width) + # Observed
  geom_hline(aes(yintercept = 1), colour = "#999999") +
  geom_crossbar(aes(y = OR, ymin = OR, ymax = 1), data = sex.prog.or %>% progs2pub(),
                fill = "#999999", colour = NA, width = dodge.width) + # Observed
  geom_point(colour = "#000000") +  # Predicted marginal means from the model
  geom_errorbar(aes(ymin = OR.F.LCL, ymax = OR.F.UCL),
                width = 0.3, colour = "#000000") +
  scale_y_continuous(limits = c(0.4, 1.21), expand = c(0, 0)) +
  labs(x = NULL, y = "Odds Ratio (Female applicants / Male applicants)") +

  theme_classic()
success.pred.sex.diff.plot



tiff("hollyplot2.tif", units="in", width=6, height=6, res=300)
success.pred.sex.diff.plot # graphics code
dev.off()
```