Assessment of potential bias in research grant peer review in Canada

Robyn Tamblyn PhD, Nadyne Girard MSc, Christina J. Qian MSc, James Hanley PhD

■ Cite as: *CMAJ* 2018 April 23;190:E489-99. doi: 10.1503/cmaj.170901

See related article at www.cmaj.ca/lookup/doi/10.1503/cmaj.180188

ABSTRACT

BACKGROUND: Peer review is used to determine what research is funded and published, yet little is known about its effectiveness, and it is suspected that there may be biases. We investigated the variability of peer review and factors influencing ratings of grant applications.

METHODS: We evaluated all grant applications submitted to the Canadian Institutes of Health Research between 2012 and 2014. The contribution of application, principal applicant and reviewer characteristics to overall application score was assessed after adjusting for the applicant's scientific productivity.

RESULTS: Among 11 624 applications, 66.2% of principal applicants were male and 64.1% were in a basic science domain. We found a significant nonlinear association between scientific productivity and final application score that differed by applicant gender and scientific domain, with higher scores associated with past funding success and *h*-index and lower scores associated with female applicants and those in the applied sciences. Significantly lower application scores were also associated with applicants who were older, evaluated by female reviewers only (v. male reviewers only, –0.05 points, 95% confidence interval [CI] –0.08 to –0.02) or reviewers in scientific domains different from the applicant's (–0.07 points, 95% CI –0.11 to –0.03). Significantly higher application scores were also associated with reviewer agreement in application score (0.23 points, 95% CI 0.20 to 0.26), the existence of reviewer conflicts (0.09 points, 95% CI 0.07 to 0.11), larger budget requests (0.01 points per $100 000, 95% CI 0.007 to 0.02), and resubmissions (0.15 points, 95% CI 0.14 to 0.17). In addition, reviewers with high expertise were more likely than those with less expertise to provide higher scores to applicants with higher past success rates (0.18 points, 95% CI 0.08 to 0.28).

INTERPRETATION: There is evidence of bias in peer review of operating grants that is of sufficient magnitude to change application scores from fundable to nonfundable. This should be addressed by training and policy changes in research funding.

Peer review is the backbone of modern science. Scientists with expertise in the field collectively make recommendations about what research is funded. Despite the almost ubiquitous use of peer review and its role in the scientific enterprise, there is limited evidence about its effectiveness.[1–3] Critics have expressed concerns about the reliability and fairness of the process, and its innate conservatism in funding interdisciplinary and innovative science.[4–6] An increasing number of empirical studies of peer review have investigated some of these criticisms. Reliability, when measured, is poor;[7–11] of greater concern is evidence suggesting the presence of systematic bias. Female scientists are less likely to be funded and published than male scientists.[12–22] Reviewers who declare conflicts of interest with an application positively bias other reviewers' rating.[23,24] Yet, few studies have taken differences in the quality of the applicant or nature of the research into account.[14,25] The first study to disentangle these effects, by adjusting for the applicant's publication impact score to quantify potential bias in the peer review of postdoctoral fellowships, reported substantial gender bias, with female scientists with the highest productivity being scored equivalent to males with the lowest productivity.[23] There have been limited attempts to separate scientific quality from potential biases in the investigator-initiated operating grant competitions that fund the bulk of science.[13,14,22,25–32] When scientific productivity is taken into account, potential gender biases are not evident in all studies, even within the same funding agency.[13,14,22,28] Differences in the characteristics of peer reviewers may explain the lack of consistency in findings, but the interaction between reviewer and applicant characteristics has not yet been investigated.

In particular, there is interest in determining whether reviewer gender, expertise, success rate, experience, scientific domain, conflict of interest and reviewer disagreement would influence and potentially bias the overall rating of an application. In this study, we used data from the national health research funding agency in Canada to estimate the reliability of peer review and to investigate potential bias in rating after differences in the scientific productivity of applicants had been taken into account.

Methods

The Canadian Institutes of Health Research (CIHR) is Canada's national health research funding agency. CIHR invests about $800 million annually in health research.[33] About 70%, or $540 million annually, is used for investigator-driven research, and $268 million for research to address strategic priorities for the country.[33] Until 2015, when reforms were made in funding programs, investigator-initiated operating grant applications were submitted to biannual competitions and were evaluated by 1 of 53 standing committees, selected as most appropriate for review by the applicant. The chair and scientific officer of each standing committee, composed of 10–15 committee members, assigned a first and second reviewer to each application, based on the committee member's self-assessment of their expertise to review each application and conflict of interest, if relevant. The first and second reviewer independently assigned preliminary scores that reflected their assessment of the quality of the application, from 1 (poor) to 4.9 (excellent); provided a written review; and presented the application and their comments to the committee for discussion. A primary and secondary reviewer consensus score was agreed to after committee discussion, then all committee members scored the application between 0.5 above or below the consensus score. The final score for an application was computed as the mean of scores assigned by all committee members, and was used to rank applications. The top-ranked applications in each committee were funded, with funded and not funded applications often differing by less than 0.1 of a point in score. If both reviewers independently assigned a score of < 3.5 to an application, the application was considered nonfundable, it was not discussed or rated by the committee, and the mean score of the 2 reviewers was used as the final score. Members who were in conflict with an application were excused during the discussion and rating.

Study population

We extracted all applications to the investigator-initiated open operating grant competition between 2012 and 2014 from the CIHR database. In this period, CIHR recorded the reviewers' self-declared expertise in reviewing each application and conflicts of interest in the central research database, which allowed their contribution to application scores to be assessed.

Variables

Application characteristics

We classified the scientific domain of the application as basic sciences (biomedical), or applied sciences (clinical, health services and policy and population health), based on the applicant's self-designation. We classified the history of the application as a new grant or a resubmission of a previously unsuccessful grant, and measured the total amount of funding requested as the sum of the amount requested per year over the grant duration.

Reviewer characteristics

We considered an application to have conflicts of interest if 1 or more members of the review panel declared a conflict of interest with the application. The self-assessed expertise of the first and second reviewer was classified as: 1) both reviewers had high expertise, 2) a mix of high and medium expertise, 3) a mix of low expertise with a high- or medium-expertise reviewer, and 4) both with low expertise. The genders of the reviewers were classified as: 1) both male, 2) male and female, 3) both female. The research experience of the first and second reviewer was based on the number of years they had applied to CIHR since its inception in 2000, and the counts for the 2 reviewers were summed to provide a continuous years-of-experience measure. Similarly, the scientific domain of the reviewer's own CIHR applications was measured as: 1) in the same scientific domain as the applicant only, 2) in mixed domains, including the domain of the applicant, 3) in domains different from that of the applicant. The proportion of applications submitted by the 2 reviewers that were successfully funded by CIHR represented the reviewers' past success at CIHR.

Applicant characteristics

The principal applicant's self-reported age, gender and primary academic institution were retrieved from the application form. When there was more than 1 principal applicant (17.5% of applications), the characteristics of the older, more senior applicant were measured.

Scientific productivity

We measured the applicant's scientific productivity by 2 indicators of academic performance and predictors of funding success: 1) previous success rate in CIHR funding, and 2) bibliometric indicators of impact. To measure CIHR funding success, we retrieved all applications submitted to CIHR since 2000 and calculated the proportion funded to provide a quantitative measure of success rate. For the bibliometric measures, we calculated the Wennerås total impact measure to enable comparisons with this study.[23] This indicator sums the impact factors of all published articles. We also calculated the *h*-index for each applicant.[34] This measure estimates the impact of a scientist's cumulative research contributions based on citations and allows an unbiased comparison of scientific achievement between individuals competing for the same resources.

To produce the bibliometric measures of scientific productivity, we used the applicant's first, middle and last name listed on the grant proposal to retrieve all publications, up to and including 2011, from the Web of Science, where the applicant was listed as an author. For each publication, we retrieved the detailed text file (authors' names, corresponding author's name and institution, publication title) and the citation reports. We found the impact factor for each journal by linking the ISSN of the journal to the Journal Citation Record file or, when there was no recorded ISSN, we used the full and abbreviated journal name to make the link.

Application scores

For each application, we retrieved the first and second reviewer scores to estimate the reliability of rating, as well as the final score to assess potential sources of systematic bias. As disagreement in ratings between reviewers may bias final application scores, we classified applications as having differences in score between the first and second reviewer of greater than 1 scale point to assess the impact of disagreement on the final application score.

Statistical analysis

We used descriptive statistics to summarize the application, applicant and reviewer characteristics. Inter-rater reliability was estimated by the intraclass correlation coefficient (ICC), where values of 0.00 to 0.40, 0.41 to 0.59, 0.60 to 0.74, and 0.75 to 1.00 are considered to represent poor, fair, good and excellent agreement, respectively.[35] We estimated the ICC for the first and second reviewer, overall and by scientific domain. Differences in within-rater variance in different scientific domains were tested using a 2-tailed *F* test.

To assess potential sources of systematic bias in rating, we estimated the association between the applicant's scientific productivity and his or her final application score using multiple linear regression within a generalized estimating equation framework to account for clustering of multiple applications from the same applicant. We used an exchangeable correlation structure to account for clustering, and added quadratic terms to assess linearity. Application was the unit of analysis, and final application score was the outcome. As the *h*-index and the total impact measure were highly correlated, we included only the *h*-index and funding success rate in the final model. We added application, reviewer and applicant characteristics to the model. In theory, after adjusting for scientific productivity, there should be no additional variance in final application score that is explained by the gender, age, reviewer expertise or agreement of the applicants. To determine whether the relationship between scientific productivity and application score was modified by applicant gender, gender mix of the reviewers, scientific domain or reviewer expertise, we included 2-way interaction terms in the model and tested using the Wald χ^2 test. As basic and applied sciences have been shown to differ in the weight given to past scientific productivity in evaluating the quality of the application,[32,36,37] we also tested the 3-way interaction between scientific productivity, science domain and applicant characteristics. To facilitate interpretation of significant interactions, we illustrated these associations graphically, and calculated the impact of these biases on final application score for common scenarios. All analyses were completed using SAS version 9.4 TS Level 1M4.

Ethics approval

This study was approved by CIHR senior executive management and the CIHR legal counsel.

Results

Overall, 11 624 applications were submitted to the open operating grant competitions between 2012 and 2014, of which 66.2% of principal applicants were male and 69.1% were aged 40 years or older (Table 1). The scientific domains of the applications were basic science (64.1%) and applied science (35.9%), of which 16.6% were clinical, 8.1% were health services and policy and 11.3% were population health. Most applications were new submissions, and more than half had 1 to 3 investigators. The mean amount of funding requested was $747 981. About 20% of applications were classified as nonfundable because both the first and second reviewer independently provided scores of less than 3.5.

In the majority of applications, both the first and second reviewer had high (16.3%) or medium-high (68.1%) expertise to review, and half had submitted their own grant applications in the same science domain in which they were reviewing (Table 1). The majority of applications were reviewed by both male and female reviewers, or male reviewers only. Most reviewers had between 10 and 20 years of combined experience, and a success rate in their own applications of between 25% and 50%. In 66.9% of applications, at least 1 member of the review panel had a conflict of interest with the application. Female applicants were more likely to apply with multiple co-investigators, ask for less funding, have their application triaged, be reviewed by female reviewers only, and have reviewers from other scientific domains.

Overall, the reliability of application rating by the first and second reviewer was fair (ICC 0.41, 95% CI 0.39 to 0.43), but only for basic science applications (ICC 0.41, 95% CI 0.39 to 0.44), whereas it was poor (ICC 0.33, 95% CI 0.30 to 0.36) for applied science applications despite greater variance between applications (Table 2). The within-rater variance component for health services and policy reviewers was almost double that of basic science reviewers (0.28 v. 0.15, $p < 0.05$).

The *h*-index and the total impact measure were highly correlated ($r = 0.8$), and showed similar trends regarding the characteristics of the application's principal investigator. Clinical investigators had the highest scientific productivity, and health services and population health researchers had the lowest. Scientific productivity was systematically lower for women and for younger applicants (Table 3). History of funding success was not strongly correlated with bibliometric measures of scientific productivity (cumulative impact: $r = 0.11$, *h*-index: $r = 0.18$), and was highest in older, male and basic science applicants. The mean final application score was highest for basic science applications.

There was a significant nonlinear association between the *h*-index, past success rate and final application score (Table 4). The greatest impact of scientific productivity on the application score was at the lower levels of the distribution. The gender and scientific domain of the applicant modified the association between past success rate and application score (significant 2- and 3-way interactions) (Figure 1). Increasing past success rate in funding had a greater positive impact on application scores in basic science compared with applied science. Overall, female applicants who had past success rates equivalent to male applicants received lower application scores, the difference being

Table 1 (part 1 of 2): Applicant and application characteristics for CIHR open operating grant applications, 2012–2014

Characteristics	All applicants, no. (%)* *n* = 11 624	Female applicants, no. (%)* *n* = 3930	Male applicants, no. (%)* *n* = 7694
Applicant (principal)			
Gender			
Female	3930 (33.8)		
Male	7694 (66.2)		
Age group, yr			
Age missing	1056 (9.1)	534 (13.6)	522 (6.8)
26–41	2968 (25.5)	1062 (27)	1906 (24.8)
42–47	2613 (22.5)	928 (23.6)	1685 (21.9)
48–54	2595 (22.3)	801 (20.4)	1794 (23.3)
55–81	2392 (20.6)	605 (15.4)	1787 (23.2)
Application			
Year submitted			
2012	4248 (36.6)	1407 (35.8)	2841 (36.9)
2013	4593 (39.5)	1577 (40.1)	3016 (39.2)
2014	2783 (23.9)	946 (24.1)	1837 (23.9)
Type of application			
New submission	7958 (68.5)	2772 (70.5)	5186 (67.4)
Resubmission	3666 (31.5)	1158 (29.5)	2508 (32.6)
No. of investigators			
1	3393 (29.2)	805 (20.5)	2588 (33.6)
2–3	3351 (28.8)	933 (23.7)	2418 (31.4)
4–5	2016 (17.3)	782 (19.9)	1234 (16.0)
> 5	2864 (24.6)	1410 (35.9)	1454 (18.9)
Initial rating above 3.4 (fundable)			
Above 3.4	9328 (80.2)	3012 (76.6)	6316 (82.1)
≤ 3.4	2296 (19.8)	918 (23.4)	1378 (17.9)
Scientific domain			
Basic science	7450 (64.1)	1858 (47.3)	5592 (72.7)
Applied science	4174 (35.9)	2072 (52.7)	2102 (27.3)
Clinical	1924 (16.6)	806 (20.5)	1118 (14.5)
Health services and policy	941 (8.1)	530 (13.5)	411 (5.3)
Population health	1309 (11.3)	736 (18.7)	573 (7.5)
Amount of funding requested, mean ± SD	$747 981 ± $445 347	$706 964 ± $504 335	$768 931 ± $410 427
Review			
Gender mix of reviewers			
Both male	5435 (46.8)	1442 (36.7)	3993 (51.9)
Both female	1459 (12.6)	742 (18.9)	717 (9.3)
1 male + 1 female	4730 (40.7)	1746 (44.4)	2984 (38.8)
Reviewer expertise			
Both high expertise	1893 (16.3)	597 (15.2)	1296 (16.8)
1 medium + 1 high or medium expertise	7917 (68.1)	2724 (69.3)	5193 (67.5)

Table 1 (part 2 of 2): Applicant and application characteristics for CIHR open operating grant applications, 2012–2014

Characteristics	All applicants, no. (%)* *n* = 11 624	Female applicants, no. (%)* *n* = 3930	Male applicants, no. (%)* *n* = 7694
1 low + 1 high, medium or low expertise	1532 (13.2)	515 (13.1)	1017 (13.2)
Both low or NA expertise	282 (2.4)	94 (2.4)	188 (2.4)
Reviewer application history			
Both from same domain as applicant	6505 (56.0)	1672 (42.5)	4833 (62.8)
Both from mixed or other domains	1842 (15.9)	939 (23.9)	903 (11.7)
1 from same domain, 1 from mixed or other domains	3277 (28.2)	1319 (33.6)	1958 (25.4)
Reviewer experience, yr			
≤ 10†	2170 (18.7)	760 (19.3)	1410 (18.3)
11–20	8842 (76.1)	2928 (74.5)	5914 (76.9)
> 20	612 (5.3)	242 (6.2)	370 (4.8)
Reviewer success rate			
No funded applications	239 (2.1)	88 (2.2)	151 (2.0)
Between 1% and < 25%	3581 (30.8)	1219 (31.0)	2362 (30.7)
Between 25% and < 50%	6655 (57.3)	2247 (57.2)	4408 (57.3)
≥ 50%	1149 (9.9)	376 (9.6)	773 (10.0)
Conflicts on the review panel			
No conflicts	3850 (33.1)	1330 (33.8)	2520 (32.8)
At least 1 conflict	7774 (66.9)	2600 (66.2)	5174 (67.2)

Note: CIHR = Canadian Institutes of Health Research, NA = not available, SD = standard deviation.
*Unless otherwise specified.
†Including no CIHR experience.

greater in applied science applications and as the past success rates increased (Figure 1). Based on the fitted model (see the note in Figure 1), a female applicant in applied sciences with a success rate of 50% would get a score of 3.75 (95% CI 3.32 to 4.18), while a male applicant would get a score of 3.82 (95% CI 3.36 to 4.28). A male applicant in applied sciences needs a funding success of 23% to get a score of 3.75 (95% CI 3.39 to 4.11). A female applicant in basic sciences with a funding success of 50% would achieve a final application score of 4.02 (95% CI 3.57 to 4.47), compared with 4.06 (95% CI 3.78 to 4.34) for males.

With regard to peer review and application characteristics, significantly lower application scores were associated with both reviewers being female (adjusted difference in score v. male reviewers only, −0.05, 95% CI −0.08 to −0.02), or the applications of both reviewers being outside of the scientific domain of the applicant (−0.07, 95% CI −0.11 to −0.03). Moreover, we observed a significant interaction between reviewer expertise and applicant past funding success, such that when both reviewers had high expertise, they were more likely to provide higher application scores to applicants with higher past success rates than were reviewers with less expertise (adjusted difference 0.18, 95% CI 0.08 to 0.29). In comparison, final application scores were higher when there was reviewer agreement (adjusted difference 0.23, 95% CI 0.20 to 0.26), a conflict with at least 1 member of the panel (0.09, 95% CI 0.07 to 0.11), for resubmissions (0.15, 95% CI 0.14 to 0.17), and for applications that requested more funding (0.01 per additional $100 000, 95% CI 0.01 to 0.02). There was no significant interaction between applicant gender and reviewer gender, or reviewer expertise.

The impact of peer review characteristics on application score is sufficient to have an impact on the likelihood of funding success. Based on the model, the estimated application score for 2 male applicants in basic science with equivalent mean scientific productivity, age and application characteristics is 3.9 for the applicant with the most favourable peer review characteristics — agreement between reviewers, conflicts on the panel, high-expertise reviewers, male reviewers only, and reviewers from the same scientific domain — compared with a score of 3.4 for the applicant without these conditions, a score that would place the application in the nonfundable range.

Interpretation

This study confirmed many of the suspected biases in the peer review of operating grant applications and identified important characteristics of peer reviewers that must be considered in application assignment. By measuring and controlling for scientific excellence of the applicant, we were able to examine how applicant, application and reviewer characteristics may unduly influence the assessment of operating grant applications. We found lower scores for applied science applications, gender inequities in application scores that favoured male applicants who had past funding success rates equivalent to female applicants, particularly in the applied sciences. Conflicts on the panel, male reviewers only, reviewers with all high expertise, and those whose own research was exclusively in the same

Table 2: Reliability* of application rating in the CIHR open operating grant competition, 2012–2014†

Scientific domain	No. of applications *n* = 11 624 (%)	Intraclass correlation	Variance components: Between applications	Variance components: Within raters
All domains	11 624 (100.0)	0.41 (0.39–0.43)	0.43	0.18
Basic science	7450 (64.1)	0.41 (0.39–0.44)	0.35	0.15
Applied science	4174 (35.9)	0.33 (0.30–0.36)	0.48	0.25
Clinical	1924 (16.6)	0.33 (0.29–0.38)	0.41	0.21
Health services and policy	941 (8.1)	0.32 (0.26–0.38)	0.54	0.28
Population health	1309 (11.3)	0.32 (0.26–0.37)	0.53	0.27

Note: CIHR = Canadian Institutes of Health Research.
*Intraclass correlation coefficient.
†Based on the first and second reviewer scores: overall, and by pillar.

Table 3: Scientific productivity in relation to the applicant characteristics*

Applicant characteristics†	Scientific productivity: *h*-index, mean ± SD	Scientific productivity: Cumulative impact,‡ mean ± SD	Scientific productivity: Historical funding success rate, 0–1;§ mean ± SD	Final application score, mean ± SD
Scientific domain				
Basic sciences	7.54 ± 4.87	142.13 ± 207.80	0.27 ± 0.21	3.87 ± 0.44
Applied sciences	6.77 ± 4.79	133.92 ± 214.06	0.22 ± 0.20	3.57 ± 0.50
Clinical	7.35 ± 4.77	161.95 ± 228.43	0.20 ± 0.20	3.63 ± 0.46
Health services and policy	6.33 ± 4.87	115.94 ± 193.11	0.22 ± 0.19	3.52 ± 0.54
Population health	6.21 ± 4.66	104.69 ± 200.23	0.23 ± 0.20	3.52 ± 0.52
Age, yr				
Missing age	6.88 ± 4.84	144.44 ± 292.00	0.24 ± 0.18	3.67 ± 0.48
26–41	5.98 ± 3.87	95.58 ± 138.60	0.20 ± 0.23	3.73 ± 0.48
42–47	7.27 ± 4.82	140.53 ± 196.01	0.25 ± 0.20	3.80 ± 0.48
48–54	7.64 ± 4.64	143.78 ± 201.39	0.26 ± 0.18	3.80 ± 0.48
55–81	8.55 ± 5.72	182.31 ± 248.42	0.31 ± 0.20	3.79 ± 0.48
Gender				
Female	5.90 ± 3.75	86.62 ± 126.70	0.23 ± 0.20	3.71 ± 0.48
Male	7.93 ± 5.19	165.00 ± 236.37	0.26 ± 0.21	3.80 ± 0.48
Overall	7.27 ± 4.86	139.22 ± 210.07	0.25 ± 0.21	3.77 ± 0.48

Note: SD = standard deviation.
*All scientific productivity analyses included 10 470 applicants who were successfully linked to the bibliometric data or 90% of our cohort.
†Principal applicant.
‡Cumulative impact is measured by summing the impact factors of all published articles (Wennerås total impact measure).
§Historical funding success rate is measured by calculating the proportion of funded applications submitted to the Canadian Institutes of Health Research since 2000.

Table 4 (part 1 of 2): Potential sources of bias in peer review of grant applications (*n* = 10 470)*

Characteristics	Mean final score	Adjusted difference in final score†	95% CI	*p* value
Applicant (principal)				
Gender by scientific domain‡				
Basic sciences				
Male	3.88	Reference		
Female	3.86	0.01	−0.06 to 0.07	0.82
Applied sciences				
Male	3.58	−0.10	−0.16 to −0.04	0.002
Female	3.56	−0.05	−0.11 to 0.01	0.13
Age group, yr§				
26–41	3.73	Reference		
42–47	3.80	−0.02	−0.05 to 0.01	0.20
48–54	3.80	−0.06	−0.09 to −0.03	< 0.0001
55–81	3.79	−0.10	−0.13 to −0.07	< 0.0001
h-index (per 1 point increase)¶	–	0.02	0.01 to 0.02	< 0.0001
h-index x *h*-index¶	–	−0.0005	−0.0007 to −0.0003	< 0.0001
h-index x gender by scientific domain¶				
Basic sciences				
Male	–	Reference		
Female	–	0.007	−0.001 to 0.014	0.08
Applied sciences				
Male	–	0.002	−0.004 to 0.008	0.44
Female	–	0.001	−0.006 to 0.007	0.78
Historical funding success (per 1% increase)	–	0.98	0.83 to 1.12	< 0.0001
Historical funding success × historical funding success¶	–	−0.55	−0.70 to −0.40	< 0.0001
Historical funding success × gender by scientific domain‡				
Basic sciences				
Male	–	Reference		
Female	–	−0.14	−0.28 to −0.01	0.04
Applied sciences				
Male	–	−0.32	−0.46 to −0.18	< 0.0001
Female	–	−0.47	−0.62 to −0.33	< 0.0001
Application				
Year submitted				
2012	3.75	Reference		
2013	3.77	0.05	0.03 to 0.07	< 0.0001
2014	3.79	0.04	0.02 to 0.06	0.0002
Type of application				
New submission	3.70	Reference		
Resubmission	3.90	0.15	0.14 to 0.17	< 0.0001
No. of investigators¶				
1	3.91			
2–3	3.77			
4–5	3.69			
> 5	3.65			
(change in score per additional investigator)	–	−0.003	−0.006 to 0.001	0.09
Amount of funding requested¶				
< $750 000	3.61			
≥ $750 000	3.91			
(change in score per $100 000 increase)	–	0.011	0.007 to 0.015	< 0.0001

Table 4 (part 2 of 2): Potential sources of bias in peer review of grant applications (*n* = 10 470)*

Characteristics	Mean final score	Adjusted difference in final score†	95% CI	*p* value
Review				
Reviewer gender mix				
Both male	3.82	Reference		
Both female	3.62	−0.05	−0.08 to −0.02	0.001
1 male + 1 female	3.75	−0.014	−0.032 to 0.003	0.11
Reviewer expertise				
Both high expertise	3.81	−0.031	−0.065 to 0.003	0.07
1 medium + 1 high or medium expertise	3.76	Reference		
1 low + 1 high, medium, or low expertise	3.76	0.003	−0.035 to 0.040	0.88
Both low or NA expertise	3.71	0.02	−0.07 to 0.10	0.70
Reviewer application history				
Both from same domain as applicant	3.87	Reference		
Both from mixed or other domains	3.67	−0.07	−0.11 to −0.03	< 0.0001
1 from same domain, 1 from mixed or other domains	3.56	−0.04	−0.07 to −0.02	< 0.0001
Reviewer experience, yr				
≤ 10	3.77	Reference		
11–20	3.77	0.002	−0.020 to 0.023	0.87
> 20	3.66	−0.02	−0.06 to 0.02	0.30
Reviewer success rate¶				
No funded applications	3.73			
Between 1% and <25%	3.73			
Between 25% and <50%	3.78			
> 50%	3.81			
(change in score per 10% increase)	–	0.056	−0.002 to 0.113	0.06
Conflicts on the review panel				
≥ 1 panel member in conflict	3.81	0.09	0.07 to 0.11	< 0.0001
No panel member in conflict	3.68	Reference		
Reviewer agreement				
Difference in score < 1	3.79	0.23	0.20 to 0.26	< 0.0001
Difference in score ≥ 1	3.49	Reference		
Applicant funding success × reviewer expertise				
Both high expertise	–	0.18	0.08 to 0.29	0.0004
1 medium + 1 high or medium expertise	–	Reference		
1 low + 1 high, medium, or low expertise	–	−0.03	−0.14 to 0.08	0.56
Both low expertise	–	−0.15	−0.44 to 0.13	0.29
Reviewer application history				
Both reviewers from same domain as applicant	3.87	Reference		
1 reviewer from same domain, 1 reviewer from mixed domains	3.56	−0.04	−0.07 to −0.02	< 0.0001
Both reviewers from mixed or other domains	3.67	−0.07	−0.11 to −0.03	< 0.0001

Note: CI = confidence interval, NA = not available.
*Association between applicant, application and reviewer characteristics and final grant application score.
†Model's intercept is 3.50.
‡We created 4 mutually exclusive categories for gender by scientific domain to facilitate interpretation of the results.
§Missing age was found to have a mean score of 3.67, and adjusted difference of −0.12 (95% CI −0.16 to −0.08, *p* < 0.0001). The estimated coefficient for age, modelled as a continuous variable, is −0.0046 (95% CI −0.0058 to −0.0033).
¶Modelled as continuous variable.

scientific domain as the applicant's conferred positive benefits in application rating.

The issue of gender inequity in peer review has been a topic of considerable debate since the original Swedish studies.[23,24] Subsequent investigations evaluated differences in success rates or application scores,[10,13–15,17,22,28,30,31,38,39] many without adjustment for scientific productivity,[22,28,30,40] an important deficiency because women have lower productivity measures. The results are mixed;[10,13–15,17,23,24,30,31,38,41] a meta-analysis suggests a modest bias of a 7% higher odds of funding success in favour of men.[17] Our results provide some possible explanation of differences across studies. We showed that the association is not linear, and is modified by scientific domain, with greater inequities for women in the applied sciences at the upper end of funding success rates. This may be why studies can show negligible to large effects depending on the scientific domain and performance of the cohort being investigated. Previous studies report that female scientists are perceived as being less competent[23,42] and having weaker leadership skills.[13,40,42] Moreover, the language used in application evaluation criteria may favour male stereotypes (e.g., "independent," "challenging").[13,15,40] In keeping with these biases, there may be greater concerns about the ability of successful female scientists to lead multiple funded projects, resulting in lower application scores, and lower funding success.

Although we did not find an interaction between the gender of the applicant and the gender mix of the reviewers, female reviewers were more stringent in their rating. Two previous studies reported similar results.[18,30] To provide equitable assessment, these systematic differences in ratings by male and female reviewers need to be addressed — for example, by reviewer training, monitoring and intervention, and possibly statistical adjustment, as is done in high-stakes professional licensing examinations.[7,43,44]

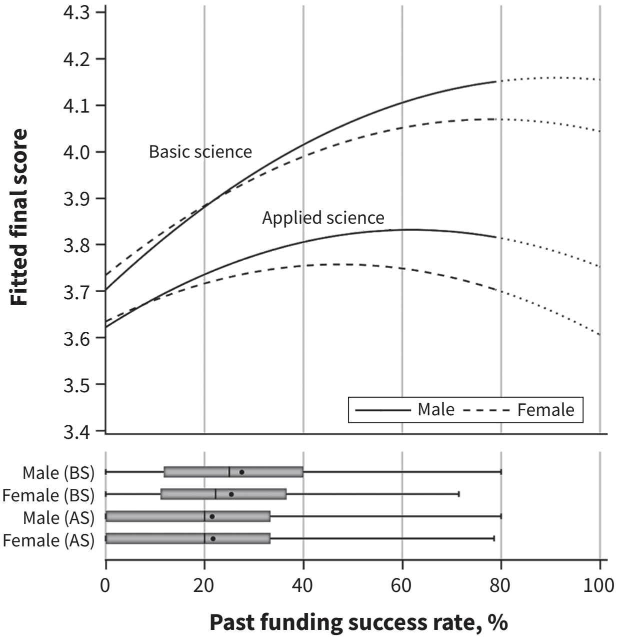

Figure 1: Fitted final scores of male and female applications in the domains of basic science (BS) and applied science (AS) in relation to past funding success rate based on the final model (Table 4). These graphs were generated using the reference category of all categorical variables in the final model, and the mean value for the continuous variables (mean number of investigators = 2, mean total amount of funding requested = $750 000, mean number of funded applications for the reviewers = 30%). For each of the subgroups, the *h*-index value was taken as the mean value within the specific group: biomedical sciences male = 7.9, biomedical sciences female = 6.3, applied sciences male = 7.9, applied sciences female = 5.5.

Our study confirmed that conflict of interest has an important positive impact on application scoring, even though panel members who have conflicts are not present for the discussion and scoring. One possible reason is that reviewers vote favourably for applicants from the same institution, even if they have never met them and would therefore not be in conflict — a phenomenon that was noted in both the French and Swedish studies.[23,24,45] Alternatively, as the same reviewers may be on the same panel for years, they may want to support the colleagues of other panel members with more positive ratings, in the spirit of collegiality. Several suggestions have been made on how to address this problem, including blinding the applicant's identity, selecting international reviewers (especially for smaller research communities), and allowing the applicant to respond to the reviewers' comments, as is done in manuscript review[45] and by some granting agencies.[18] To date, there is no evidence on whether these strategies mitigate conflict bias in peer review.

Our analyses provide novel evidence about the effect of reviewer expertise and the scientific domain of their own applications on application rating. Of particular interest was the observation that high-expertise reviewers were more likely to pay attention to the applicant's past funding success rate, rating the applications from more successful scientists higher. There has been very limited exploration of reviewer expertise and the role it plays in grant review. A recent study of reviewer expertise at the National Institutes of Health suggests that reviewers with higher levels of expertise are more informed and positively biased in their rating of projects in their own area.[29] These preliminary results suggest that the reviewers' own grant and publication track record, as well as their self-reported expertise, should be considered in reviewer assignments.

When combined, reviewer characteristics can have a substantial effect on an application's score and its likelihood of funding. In the worst-case scenario, an applicant who has female reviewers only, no conflicts on the committee, disagreement in the quality of the application by the reviewers, and reviewers with less expertise in the domain may receive a score 0.5 points lower on a 1 to 4.9 scale. A difference of this size could move an application with a fundable score of 3.9 to a nonfundable score of 3.4. Future research should be directed toward better methods of matching reviewers to applications, and monitoring and correcting for potential reviewer biases.

Similar to many other studies,[7–10] we found that the reliability of scientific review was fair to poor. Moreover, we found that disagreement between reviewers systematically lowered the score of an application. Increasing the number of reviewers has been recommended as an effective means of improving reliability.[46] Also, as noted in another study, reviewers give different weights to evaluation criteria such as originality, usefulness, methodology and feasibility.[47] Structuring and rating each component is recommended to address this problem, by providing explicit, transparent weighting of assessment.[47] In addition, training has been shown to be effective in getting reviewers to use rating scales in the same way.[7]

Limitations

There are important limitations to consider in the interpretation of the results. Although we used standard measures to assess the scientific excellence of the applicant, we had no external gold standard measure of the quality of the proposal. The improvement in scoring seen with resubmissions, and with higher funding requests, which have been reported previously,[15,25,48] may represent true superiority in the quality of the proposal; however, it is unlikely that biases related to reviewer characteristics or scientific domain are related to differences in proposal quality. We were conservative in our linkage of applicants to publications, requiring perfect agreement on first and last name. Our approach likely underestimated the bibliometric measures of productivity and impact, possibly differentially penalizing female scientists if they changed their name after marriage. Finally, there may be other factors that influence application score that we could not measure, such as the quality of the institution or department.

Conclusion

We identified potential systematic biases in peer review that penalize female applicants and are associated with peer reviewer characteristics; these may be addressed through policy change, training and monitoring.

References

1. Demicheli V, Di Pietrantoni C. Peer review for improving the quality of grant applications. *Cochrane Database Syst Rev* 2007;(2):MR000003.
2. Mayo NE, Brophy J, Goldberg MS, et al. Peering at peer review revealed high degree of chance associated with funding of grant applications. *J Clin Epidemiol* 2006;59:842.
3. Graves N, Barnett AG, Clarke P. Funding grant proposals for scientific research: retrospective analysis of scores by members of grant review panel. *BMJ* 2011; 343:d4797.
4. Rennie D. Let's make peer review scientific. *Nature* 2016;535:31-3.
5. Mervis J. Peering into peer review. *Science* 2014;343:596-8.
6. Rabesandratana T. The seer of science publishing. *Science* 2013;342:66-7.
7. Sattler DN, McKnight PE, Naney L, et al. Grant peer review: improving inter-rater reliability with training. *PLoS One* 2015;10:e0130450.
8. Marsh HW, Jayasinghe UW, Bond N. Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. *Am Psychol* 2008;63:160-8.
9. Mutz R, Bornmann L, Daniel HD. Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: a general estimating equations approach. *PLoS One* 2012;7:e48509.
10. Jayasinghe UW, Marsh HW, Bond N. A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *J R Stat Soc Series B Stat Methodol* 2003;166:279-300.
11. Pier EL, Raclaw J, Kaatz A, et al. 'Your comments are meaner than your score': score calibration talk influences intra- and inter-panel variability during scientific grant peer review. *Res Eval* 2017;26:1-14.
12. Sheridan J, Savoy JN, Kaatz A, et al. Write more articles, get more grants: the impact of department climate on faculty research productivity. *J Womens Health (Larchmt)* 2017;26:587-96.
13. 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 (Larchmt)* 2017;26:560-70.
14. Ginther DK, Kahn S, Schaffer WT. Gender, race/ethnicity, and National Institutes of Health R01 research awards: Is there evidence of a double bind for women of color? *Acad Med* 2016;91:1098-107.
15. 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.
16. Jagsi R, DeCastro R, Griffith KA, et al. Similarities and differences in the career trajectories of male and female career development award recipients. *Acad Med* 2011;86:1415-21.
17. Bornmann L, Mutz R, Daniel H-D. Gender differences in grant peer review: a meta-analysis. *J Informetrics* 2007;1:226-38.
18. Jayasinghe UW, Marsh HW, Bond N. Peer review in the funding of research in higher education: the Australian experience. *Educ Eval Policy Anal* 2001;23:343-64.
19. Larivière V, Vignola-Gagné E, Villeneuve C, et al. Sex differences in research funding, productivity and impact: an analysis of Québec university professors. *Scientometrics* 2011;87:483-98.
20. Gannon F, Quirk S, Guest S. Searching for discrimination: Are women treated fairly in the EMBO postdoctoral fellowship scheme? *EMBO Rep* 2001;2:655-7.
21. Symonds MR, Gemmell NJ, Braisher TL, et al. Gender differences in publication output: towards an unbiased metric of research performance. *PLoS One* 2006;1:e127.
22. Kaatz A, Lee YG, Potvien A, et al. Analysis of National Institutes of Health R01 application critiques, impact, and criteria scores: Does the sex of the principal investigator make a difference? *Acad Med* 2016;91:1080-8.
23. Wennerås C, Wold A. Nepotism and sexism in peer-review. *Nature* 1997;387:341-3.
24. Sandström U, Hällsten M. Persistent nepotism in peer-review. *Scientometrics* 2008;74:175-89.
25. Eblen MK, Wagner RM, RoyChowdhury D, et al. How criterion scores predict the overall impact score and funding outcomes for National Institutes of Health peer-reviewed applications. *PLoS One* 2016;11:e0155060.
26. Ginther DK, Schaffer WT, Schnell J, et al. Race, ethnicity, and NIH research awards. *Science* 2011;333:1015-9.
27. Ginther DK, Haak LL, Schaffer WT, et al. Are race, ethnicity, and medical school affiliation associated with NIH R01 type 1 award probability for physician investigators? *Acad Med* 2012;87:1516-24.
28. Pohlhaus JR, Jiang H, Wagner RM, et al. Sex differences in application, success, and funding rates for NIH extramural programs. *Acad Med* 2011;86:759-67.
29. Li D. Expertise versus bias in evaluation: evidence from the NIH. *Am Econ J Appl Econ* 2017;9:60-92.
30. Mutz R, Bornmann L, Daniel H-D. Does gender matter in grant peer review? An empirical investigation using the example of the Austrian science fund. *Z Psychol* 2012;220:121-9.
31. *Women and peer review: an audit of the Wellcome Trust's decision-making on grants*. Report No 8. London (UK): The Wellcome Trust; 1997.
32. Tamblyn R, McMahon M, Girard N, et al. Health services and policy research in the first decade at the Canadian Institutes of Health Research. *CMAJ Open* 2016;4:E213-21.
33. *CIHR internal assessment — report for the 2011 international review. Part 2: CIHR's budget*. Ottawa: Canadian Institutes of Health Research; 2011.
34. Hirsch JE. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci U S A* 2005;102:16569-72.
35. Cicchetti D. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 1994;6:284-90.
36. Nicol MB, Henadeera K, Butler L. NHMRC grant applications: a comparison of "track record" scores allocated by grant assessors with bibliometric analysis of publications. *Med J Aust* 2007;187:348-52.

37. Kingwell BA, Anderson GP, Duckett SJ, et al.; National Health and Medical Research Council Evaluations and Outcomes Working Committee. Evaluation of NHMRC funded research completed in 1992, 1997 and 2003: gains in knowledge, health and wealth. *Med J Aust* 2006;184:282-6.

38. Bornmann L, Mutz R, Daniel H-D. How to detect indications of potential sources of bias in peer review: a generalized latent variable modeling approach exemplified by a gender study. *J Informetrics* 2008;2:280-7.

39. Head MG, Fitchett JR, Cooke MK, et al. 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.

40. Kaatz A, Magua W, Zimmerman DR, et al. A quantitative linguistic analysis of National Institutes of Health R01 application critiques from investigators at one institution. *Acad Med* 2015;90:69-75.

41. Ward JE, Donnelly N. Is there gender bias in research fellowships awarded by the NHMRC? *Med J Aust* 1998;169:623-4.

42. Moss-Racusin CA, Dovidio JF, Brescoll VL, et al. Science faculty's subtle gender biases favor male students. *Proc Natl Acad Sci U S A* 2012;109:16474-9.

43. Abrahamowicz M, Tamblyn RM, Ramsay JO, et al. Detecting and correcting for rater-induced differences in standardized patient tests of clinical competence. *Acad Med* 1990;65(Suppl):S25-6.

44. Engelhard G Jr, Myford CM. *Monitoring faculty consultant performance in the advanced placement English literature and composition program with a many-faceted Rasch model*. Report No. 2003-1. College Board Research Report No 2003-1 (ETS RR-03-01). New York: College Entrance Examination Board; 2003.

45. Abdoul H, Perrey C, Tubach F, et al. Non-financial conflicts of interest in academic grant evaluation: a qualitative study of multiple stakeholders in France. *PLoS One* 2012;7:e35247.

46. Snell RR. Menage a quoi? Optimal number of peer reviewers. *PLoS One* 2015;10:e0120838.

47. Abdoul H, Perrey C, Amiel P, et al. Peer review of grant applications: criteria used and qualitative study of reviewer practices. *PLoS One* 2012;7:e46054.

48. *Success rates for grant funding*. Research Councils UK; 2014.

Competing interests: None declared.

This article has been peer reviewed.

Affiliations: Institute of Health Services and Policy Research (Tamblyn), Canadian Institutes of Health Research, Ottawa, Ont.; Departments of Medicine (Tamblyn) and Epidemiology, and Biostatistics (Girard, Qian, Hanley), Faculty of Medicine, McGill University, Montréal, Que.

Contributors: All authors contributed substantially to study design, analysis, and interpretation of data. All authors also participated in drafting and revising the manuscript, providing final approval of the version to be published, and agree to be accountable for all aspects of the work.

Funding: Funding was provided by the Canadian Institutes of Health Research (CIHR). The study sponsor approved the use of the data, and the original manuscript as per CIHR policy. CIHR had no role in the design of the study, the analysis of interpretation of the data, or the the review and approval of the final published manuscript.

Acknowledgements: The authors wish to acknowledge David Peckham and his team from the Canadian Institutes of Health Research for assembling the data for this study and answering many questions; and Dr. Jeff Latimer for reviewing and providing feedback on the manuscript.

Accepted: Nov. 23, 2017

Correspondence to: Robyn Tamblyn, robyn.tamblyn@mcgill.ca