

Washington University School of Medicine

Digital Commons@Becker

2020-Current year OA Pubs

Open Access Publications

8-8-2023

Influence of participant and reviewer characteristics in application scores for a hematology research training program

Sara K Vesely

Allison King

Emily Vettese

John G Heller

Adam Cuker

See next page for additional authors

Follow this and additional works at: https://digitalcommons.wustl.edu/oa_4



Part of the [Medicine and Health Sciences Commons](#)

Please let us know how this document benefits you.

Authors

Sara K Vesely, Allison King, Emily Vettese, John G Heller, Adam Cuker, Cecelia Calhoun, Wendy Stock, Morgan Homer, Josel Fritz, and Lillian Sung

Influence of participant and reviewer characteristics in application scores for a hematology research training program

Sara K. Vesely,¹ Allison King,² Emily Vettese,³ John G. Heller,¹ Adam Cuker,⁴ Cecelia Calhoun,⁵ Wendy Stock,⁶ Morgan Homer,⁷ Josel Fritz,⁷ and Lillian Sung^{3,8}

¹Department of Biostatistics and Epidemiology, Hudson College of Public Health, University of Oklahoma Health Sciences Center, Oklahoma City, OK; ²Division of Pediatric Hematology/Oncology and Division of Public Health Sciences in Department of Surgery, Washington University School of Medicine Program in Occupational Therapy, St. Louis, MO; ³Program in Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, ON; ⁴Department of Medicine and Department of Pathology & Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; ⁵Section of Hematology, Department of Internal Medicine, Yale University School of Medicine, New Haven, CT; ⁶Department of Medicine, University of Chicago Comprehensive Cancer Center, Chicago, IL; ⁷American Society of Hematology, Washington, DC; and ⁸The Division of Haematology/Oncology, The Hospital for Sick Children, Toronto, ON

Key Points

- Among American Society of Hematology CRTI applicants, URM applicants received significantly lower scores than non-URM applicants.
- Impact of the reviewer's sex and URM status on application scores changed over time.

The American Society of Hematology Clinical Research Training Institute (CRTI) is a clinical research training program with a competitive application process. The objectives were to compare application scores based on applicant and reviewer sex and underrepresented minority (URM) status. We included applications to CRTI from 2003 to 2019. The application scores were transformed into a scale from 0 to 100 (100 was the strongest). The factors considered were applicant and reviewer sex and URM status. We evaluated whether there was an interaction between the characteristics and time related to application scores. In total, 713 applicants and 2106 reviews were included. There was no significant difference in scores according to applicant sex. URM applicants had significantly worse scores than non-URM applicants (mean [standard error] 67.9 [1.56] vs 71.4 [0.63]; $P = .0355$). There were significant interactions between reviewer sex and time ($P = .0030$) and reviewer URM status and time ($P = .0424$); thus, results were stratified by time. For the 2 earlier time periods, male reviewers gave significantly worse scores than did female reviewers; this difference did not persist for the most recent time period. The URM reviewers did not give significantly different scores across time periods. URM applicants received significantly lower scores than non-URM applicants. The impact of reviewer sex and URM status changed over time. Although male reviewers gave lower scores in the early periods, this effect did not persist in the late period. Efforts are required to mitigate the impact of applicant URM status on application scores.

Introduction

The American Society of Hematology (ASH) has been dedicated to improving the conduct of patient-oriented research for almost 2 decades and has begun this effort by developing and sustaining a training program called the Clinical Research Training Institute (CRTI).¹ This program started in North

Submitted 20 January 2023; accepted 10 March 2023; prepublished online on *Blood Advances* First Edition 20 March 2023; final version published online 28 July 2023. <https://doi.org/10.1182/bloodadvances.2023009792>.

Original data are available on request from the corresponding author, Lillian Sung (lillian.sung@sickkids.ca).

The full-text version of this article contains a data supplement.

© 2023 by The American Society of Hematology. Licensed under [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International \(CC BY-NC-ND 4.0\)](https://creativecommons.org/licenses/by-nc-nd/4.0/), permitting only noncommercial, nonderivative use with attribution. All other rights reserved.

America in 2003 and restricts attendance to senior fellows and junior faculty focused on classical or malignant hematology research.^{2,3} Typically, the CRTI is a 1 year program that consists of in-person and remote interactive sessions. This includes didactic learning, workshops, protocol development, and mentorship.

Since 2016, the program has been guided by a steering subcommittee with 4 specific foci, namely disparity reduction, curriculum development, evaluation, and mentorship. Efforts to reduce disparities, including race, ethnicity, sex, and socioeconomic status, have been a core component of program evolution. There are many opportunities for disparities to disadvantage CRTI participation, including lack of awareness of the CRTI program, failure to submit an application, absence of mentorship to develop a strong proposal, and lack of resources to undertake additional training.

Unconscious bias has been increasingly recognized as a barrier to academic success.⁴⁻⁶ In the CRTI program, a time point at which unconscious bias could be important is during the selection of applicants to participate in the program. Unconscious bias may be related to how the reviewer views applicant characteristics or the reviewer's own characteristics. Despite the potential for unconscious bias, little is known about whether this issue contributes to the applicant selection for clinical research training. We focused on underrepresented minority (URM) status, defined as Blacks or African Americans, Hispanics or Latinos, American Indians or Alaska Natives, or Native Hawaiians and other Pacific Islanders, because they have been shown to be underrepresented in biomedical research.⁷

We hypothesized that applicant and reviewer characteristics such as sex and URM status could affect the acceptance to CRTI. Consequently, the primary objective was to compare application scores based on the applicant's sex and URM status. The secondary objective was to determine whether reviewer attributes contributed to application scores and whether the effect of reviewer attributes differed based on applicant attributes.

Methods

CRTI program

We have previously described the CRTI program in depth.^{8,9} In short, CRTI is a mentored training program that focuses on protocol development, clinical research education, and networking opportunities. From 2003 to 2019, applicants were senior fellows or junior faculty members within the first 3 years of their first faculty appointment with a planned career in patient-oriented hematology research. Most participants resided in the United States or Canada, although international applicants were also eligible. Except for the initial 3 years, the program had a 1 year duration and included a weeklong summer workshop held in August and 2 in-person meetings in the following December and May. The summer workshop consisted of didactic sessions, interactive workshops, small groups focused on protocol development, and opportunities for interactions with other participants and faculty members. Faculty members were established researchers in patient-oriented investigations, biostatisticians, and representatives from key funding agencies such as the National Institutes of Health. Starting in 2011, the trainees were matched to a CRTI mentor with minimum

quarterly contact throughout the 1 year program. Proposals can focus on adult and pediatric populations.

Application content and review

After a letter-of-intent stage, eligible applicants were invited to submit full proposals in March for the program starting in August. The full application consisted of the application form, demographic survey, career development plan, research proposal, the applicant's home (institutional) mentor's biosketch, the home mentor's letter of support, and an institutional commitment letter from a division chief or a similar institutional official. The demographic survey included questions regarding the applicant's sex, race, and ethnicity but allowed the participants to leave these questions blank.

The study section to assess the full proposals was held in May each year. The reviewers were ASH members who were clinical researchers; they were selected by the program's senior and junior co-director at that time. Each year, between 20 and 30 reviewers were selected, and they received written guidance on how to score applications. Every application was assigned to a primary, secondary, or tertiary reviewer; each submitted an overall score and critique of the application. Reviewers were asked to consider the research proposal, potential of the applicant (based on the biosketch and career development plan), home mentor biosketch, and institutional commitment letter. The research proposal was scored based on its significance, approach, feasibility, and innovation. Each year, 20 applicants were chosen to participate in CRTI, although up to 3 additional applicants could be selected to promote diversity.

The study section was held remotely in the first 2 years of the program but then transitioned to an in-person meeting at ASH headquarters in Washington, DC, between 2005 and 2019. During the study section, the strongest and weakest scores were accepted and triaged by the co-directors, respectively; reviewers were offered the ability to discuss any of these applications. Among the applications to be discussed, the primary, secondary, and tertiary reviewers announced their original scores. The primary reviewer then summarized the application and its strengths and weaknesses. The secondary and tertiary reviewers added their comments. The application was then opened for discussion among all reviewers. After the discussion, the primary, secondary, and tertiary reviewers announced their revised scores, and the entire study section silently scored the application. The final selection of accepted applications considered the study section average or median score and diversity based on sex, URM status, classical vs malignant hematology, adult vs pediatric focus, and the institution or program. The applicants self-reported their race and ethnicity in the applications evaluated by the reviewers. URM status was defined as the applicant self-reporting 1 of the following: (1) racial background of Black or African American, American Indian or Alaskan native, or native Hawaiian or other Pacific Islander or (2) Hispanic ethnicity.

Study population

Application reviews for the cohorts from 2003 to 2019 were included; application reviews beyond 2019 were not included because the procedures were modified in 2020 because of the COVID-19 pandemic. The records of each study section were

maintained differently throughout this period, and no scores were retrievable for 2007, 2009, or 2014. For 2010, only the study section average scores were available, not the individual review scores; thus, 2010 was also excluded. Thus, eligible applicants and reviews were those for the cohorts from 2003 to 2006, 2008, from 2011 to 2013, and from 2015 to 2019. When applicants in eligible years could not be uniquely identified (in some years, some records only included initials), these applications and their associated reviews were excluded.

Outcomes and exposure variables

The primary outcome was the individual primary, secondary, and tertiary reviewer scores. The specific scoring systems have changed over time and are outlined in [Appendix 1](#). From 2003 to 2006, the scoring rubric ranged from 1 to 10, in which 10 was considered the strongest application. In 2008, the scoring rubric ranged from 1 to 15, in which 15 was considered the strongest application. From 2011 to 2019, the direction of scoring was reversed; the lowest score was considered the strongest application, with a scoring ranging from 3 to 15 in 2011 and from 4 to 36 between 2012 and 2019. The scores were transformed into a common scale that ranged from 0 to 100, in which 100 was the strongest application possible.

The factors considered were applicant and reviewer characteristics. For applicants, sex, URM status, race, and ethnicity were evaluated. For the reviewers, sex and URM status were evaluated.

Statistical analysis

The study section years were categorized into 3 time periods to keep the number of years similar while minimizing the skipped years within the time periods: from 2003 to 2006; from 2008 to 2013 (including 2008 and 2011-2013); and from 2015 to 2019.

The demographic characteristics of applicants and reviewers were compared based on the time period using χ^2 test or Fisher exact test. To compare mean application scores based on time period, we created mixed models that accounted for the correlation of scores based on the applicant (using their ASH identification number) within a study section year. For applicants who applied for multiple years, only correlation by multiple reviewers for a given year was taken into account and not the correlation across different years.

To evaluate whether applicant or reviewer characteristics were associated with application scores, multivariate mixed models were created using compound symmetry as the covariance structure and random intercepts for each applicant's unique ASH number and study section year. Each model accounted for the time period and interaction between the time period and the characteristics under investigation. If the interaction was significant (suggesting that the effect of the characteristic on application scores changed over time), then the effect of that characteristic was determined separately for each time period. To evaluate whether the effect of reviewer characteristics differed based on applicant characteristics, an interaction term was added to the model and specifically examined.

All tests were 2-sided, and statistical significance was defined as $P < .05$. The analysis was conducted using R, a language and

environment for statistical computing (The R Foundation for Statistical Computing, Vienna, Austria) and SAS 9.4 (Cary, NC).

Results

Among the eligible applicants and reviews, 713 applicants and 2106 reviews were included in the analysis. There were 537 unique applicants, with 71 individuals applying multiple times. More specifically, 67 applied 2 times (39 were accepted the second time); 3 applied 3 times (1 was accepted the third time) and 1 applied 4 times (not accepted). [Figure 1](#) shows the flow diagram of applicants and reviews, including the reasons for exclusion. The numbers of applicants in the study section eras were as follows: from 2003 to 2006 ($n = 168$), 2008 or from 2011 to 2013 ($n = 204$), and from 2015 to 2019 ($n = 341$). [Table 1](#) illustrates the demographic characteristics of the applicants and those who were accepted to CRTI based on the time period. Over time, there was significantly more diversity in applicants based on the URM status, race, and ethnicity. [Table 2](#) illustrates the demographic characteristics of the reviewers based on the time period. Over time, there was significantly more diversity among reviewers based on sex, URM status, race, and ethnicity, noting that if a reviewer evaluated 7 grants at a study section, they were counted 7 times.

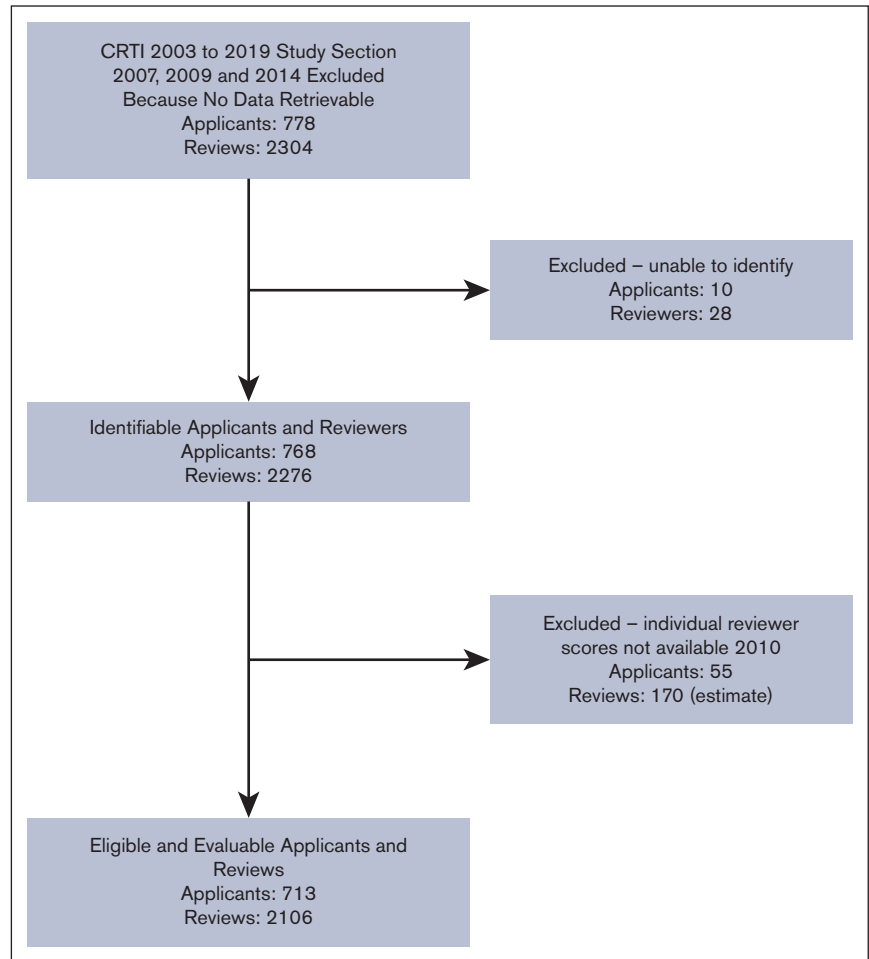
When evaluating the impact of applicant characteristics on application scores, there were no significant interactions based on characteristics and time period (data not shown); consequently, effects were observed across all applicants. [Table 3](#) shows the mean of the initial scores of the primary, secondary, and tertiary reviewers based on applicant characteristics. There was no significant difference in the scores based on applicant sex. URM applicants had significantly worse scores on average than non-URM applicants (mean score [standard error], 67.9 [1.6] vs 71.4 [0.6]; $P = .0355$). Hispanic applicants also had lower mean scores than non-Hispanic applicants (67.0 [2.1] vs 71.3 [0.6]; $P = .0453$).

When evaluating the impact of reviewer characteristics on application scores, there were significant interactions between characteristics and time for sex ($P = .0030$) and URM status ($P = .0424$). Thus, the scores were presented separately for each time period and are shown in [Table 4](#). For the 2 earlier time periods, male reviewers gave significantly worse scores than female reviewers; this difference did not persist for the most recent time period. URM reviewers did not give significantly different scores compared with non-URM reviewers for any of the 3 time periods. [Table 4](#) also shows the interactions between reviewer and applicant sex and reviewer and applicant URM status. There was no significant interaction based on sex in any of the 3 time periods. Similarly, the effect of a URM reviewer did not differ based on the URM status of the applicant for time periods 1 and 3 ($P_{\text{interaction}} = .2082$ and $.2295$). For time period 2, the interaction P value was $.0104$, indicating that URM reviewers scored URM applicants higher than non-URM applicants, whereas non-URM reviewers scored non-URM applicants higher than URM applicants.

Discussion

In this evaluation of CRTI applications over a 17-year period, we found that diversity among applicants based on URM status, race, and ethnicity and that among reviewers based on sex, URM status, race, and ethnicity increased over time. URM applicants had

Figure 1. Flow of applicants and reviews in the study.



significantly lower scores than non-URM applicants. We also found that the disparity in reviewer sex scores changed over time, with male reviewers giving significantly lower scores than female reviewers during the first 2 time periods but not the most recent time period.

We showed that application scores were significantly lower when applicants were URM or Hispanic. These effects might have been influenced by confounders, including the environment, previous research experience, and mentorship. However, unconscious bias is possible. Unconscious bias is important because it is pervasive and may be more prevalent than overt forms of bias.¹⁰ Disparity in successful grant applications has been observed for applicants who are Black or African American.¹¹ Our ability to analyze race and ethnicity based on the applicant and reviewer is important because previous efforts to evaluate these characteristics were limited owing to the lack of availability of these data elements.^{10,12}

We did not find that the application scores for female applicants differed from those for male applicants. In contrast to our finding, bias based on sex in academic settings has been observed during grant peer review.^{12,13} In addition, manuscripts with female first authors received significantly lower scores in peer review and were cited less often compared with those with male first authors.¹⁴ Although we found that women and men had similar application scores, our other research has demonstrated that female CRTI

alumnae experienced less academic success, as measured by publications and protected time for research.^{8,9}

All analyses stratified based on the time period should be considered to be hypothesis generating. We took this approach because we observed significant interactions between reviewer characteristics and time period. However, major concerns with this approach include multiple tests and smaller sample sizes for each comparison, possibly leading to both false-positive and false-negative results. Specifically, the significant interaction between the reviewer URM status and applicant URM status during the second time period should be viewed cautiously, given these concerns and the small number of URM reviewers and applicants ($n = 3$).

If a review of CRTI applications is influenced by applicant characteristics, such as URM status, what action should arise from these findings? Although we cannot rule out confounders that could explain our findings, they suggest that active approaches to mitigate the potential for unconscious bias are warranted. One change that has already been made in response to this analysis was the modification of the application scoring rubric to add an overall priority score. This addition encourages a more holistic application review and prompts reviewers to consider diversity and the "distance traveled" as a component of the overall priority score. Other approaches include an explicit discussion of our findings and calibration exercises before the review process. Another option could be concealing the applicant

Table 1. Demographics of CRTI applicants (N = 713) and accepted applicants (n = 265)

| Study section year | 2003-2006 | 2008, 2011, 2012, 2013 | 2015-2019 | P value* |
|-------------------------------------|---------------|------------------------|----------------|----------|
| | n = 168 | n = 204 | n = 341 | |
| All applicants, n (%) | | | | |
| Sex | | | | .3148 |
| Male | 81 (48.2) | 84 (41.2) | 160 (46.9) | |
| Female | 87 (51.8) | 120 (58.2) | 181 (53.1) | |
| Underrepresented minority | | | | < .001 |
| Yes | 15 (8.9) | 20 (9.8) | 46 (13.5) | |
| No | 98 (58.3) | 149 (73.0) | 247 (72.4) | |
| Unknown | 55 (32.7) | 35 (17.2) | 48 (14.1) | |
| Race | | | | < .001 |
| Black or African American | 5 (3.0) | 8 (3.9) | 22 (6.5) | |
| American Indian or Alaskan native | 2 (1.2) | 0 | 1 (0.3) | |
| White | 73 (43.5) | 104 (51.0) | 161 (47.2) | |
| Asian | 26 (15.5) | 46 (22.6) | 92 (27.0) | |
| Other | 8 (4.8) | 19 (9.3) | 16 (4.7) | |
| Unknown | 54 (32.1) | 27 (13.2) | 49 (14.4) | |
| Ethnicity | | | | <.001 |
| Hispanic | 8 (4.8) | 12 (5.9) | 25 (7.3) | |
| Not Hispanic | 104 (61.9) | 161 (78.9) | 298 (87.4) | |
| Unknown | 56 (33.3) | 31 (15.2) | 18 (5.3) | |
| Accepted to CRTI | | | | |
| Yes | 79 (47.0) | 81 (39.7) | 105 (30.8) | .0012 |
| No | 89 (53.0) | 123 (60.3) | 236 (69.2) | |
| Attended CRTI | | | | .0009 |
| Yes | 79 (47.0) | 81 (39.7) | 104 (30.5) | |
| No | 89 (53.0) | 123 (60.3) | 237 (69.5) | |
| Accepted to CRTI, n (%) | N = 79 | N = 81 | N = 105 | |
| Position at CRTI | | | | .0409 |
| Fellow | 55 (69.6) | 48 (59.3) | 50 (47.6) | |
| Faculty | 19 (24.1) | 22 (27.2) | 38 (36.2) | |
| Other | 5 (6.3) | 11 (13.6) | 17 (16.2) | |
| Medicine or pediatric focus at CRTI | | | | .3384 |
| Adult | 49 (62.0) | 54 (66.7) | 78 (74.3) | |
| Pediatric | 24 (30.4) | 24 (29.6) | 24 (22.9) | |
| Both | 6 (7.6) | 3 (3.7) | 3 (2.9) | |
| Clinical focus at CRTI | | | | .0002 |
| Malignant hematology | 34 (43.0) | 47 (58.0) | 66 (62.9) | |
| Benign hematology | 28 (35.4) | 27 (33.3) | 38 (36.2) | |
| Both | 6 (7.6) | 2 (2.5) | 1 (0.9) | |
| Unknown | 11 (13.9) | 5 (6.2) | 0 (0.0) | |

*P value calculated using χ^2 or Fisher Exact test.

demographic characteristics, including URM status, from the reviewers. Future work should examine inequities apart from study section scores that may influence research success.

Given the challenges in obtaining individual reviewer scores for CRTI, we focused on the short-term outcomes of CRTI acceptance. Although understanding the disparities in CRTI application scores is

important, we did not evaluate long-term academic outcomes, which are more salient. It is important for future research to build upon this work to examine how sex, URM status, and reviewer scores ultimately affect academic success. More specifically, it is important to evaluate the features that predict the success of clinician scientists or clinician leaders 2, 5, or 10 years after CRTI completion.

Table 2. Demographic characteristics of CRTI reviewers by sex and underrepresented minority status and mean scores (N = 2106 reviews)

| Study section year | 2003-2006 | 2008, 2011, 2012, 2013 | 2015-2019 | P value† |
|---|---------------|------------------------|----------------|----------|
| | n = 477 n (%) | n = 611 n (%) | n = 1018 n (%) | |
| Reviewer sex | | | | < .001 |
| Male | 284 (59.5) | 188 (30.8) | 311 (30.5) | |
| Female | 132 (27.7) | 317 (51.9) | 631 (62.0) | |
| Unknown | 61 (12.8) | 106 (17.3) | 76 (7.5) | |
| Reviewer underrepresented minority | | | | < .001 |
| Yes | 11 (2.3) | 37 (6.1) | 239 (23.4) | |
| No | 260 (54.5) | 437 (71.5) | 693 (68.3) | |
| Unknown | 206 (43.2) | 137 (22.4) | 86 (8.3) | |
| Reviewer race | | | | < .001 |
| Black or African American | 11 (2.3) | 18 (2.9) | 124 (12.2) | |
| American Indian or Alaskan native | 0 | 7 (1.1) | 0 | |
| White | 353 (74.0) | 345 (56.5) | 602 (59.1) | |
| Asian | 0 | 93 (15.2) | 182 (17.9) | |
| Other | 0 | 36 (5.9) | 24 (2.4) | |
| Unknown | 113 (23.7) | 112 (18.3) | 86 (8.4) | |
| Reviewer ethnicity | | | | < .001 |
| Hispanic | 0 | 12 (2.0) | 115 (11.3) | |
| Not Hispanic | 271 (56.8) | 462 (75.6) | 818 (80.4) | |
| Unknown | 206 (43.2) | 137 (22.4) | 85 (8.3) | |
| Mean score* (standard error) | 68.4 (1.1) | 67.0 (1.0) | 71.3 (0.8) | .0017 |

*Least squares means from a mixed model taking into account the correlation of scores per applicant within a study section year.

†Each application's reviewed by a primary, secondary, or tertiary reviewer. P values were compared using the χ^2 or Fisher exact test.

The strength of this report is its ability to evaluate applicant and reviewer characteristics over a lengthy period of time. However, limitations exist. Heterogeneous scoring mechanisms were used

over time, but the scores were transformed to facilitate comparisons. We lacked data on factors such as socioeconomic status, which would have strengthened the analysis.

Table 3. Applicant review scores and impact of applicant sex and underrepresented minority status (N = 2106)

| Characteristic | n | Mean scores* | Standard error | P value† |
|-----------------------------------|------|--------------|----------------|----------|
| Sex applicant | | | | .1888 |
| Male | 963 | 68.6 | 0.7 | |
| Female | 1143 | 70.0 | 0.7 | |
| URM applicant‡ | | | | .0355 |
| Yes | 241 | 67.9 | 1.6 | |
| No | 1456 | 71.4 | 0.6 | |
| Race applicant‡ | | | | .1685 |
| Black or African American | 104 | 68.5 | 2.3 | |
| American Indian or Alaskan native | 9 | 75.6 | 8.0 | |
| White | 996 | 72.0 | 0.8 | |
| Asian | 486 | 69.8 | 1.1 | |
| Other | 126 | 68.0 | 2.1 | |
| Ethnicity applicant‡ | | | | .0453* |
| Hispanic | 134 | 67.0 | 2.1 | |
| Not Hispanic | 1662 | 71.3 | 0.6 | |

*Least squares means from a mixed model taking into account the correlation of scores per applicant within a study section year.

†P value controlling for time period and excluding missing/unknown characteristics.

‡Unknown for URM (n = 409), race (n = 385), and ethnicity (n = 310).

Table 4. Applicant review scores and impact of reviewer and applicant characteristics (N = 2 106) based on the study section year

| Study section year | 2003-2006 | | | 2008, 2011, 2012, 2013 | | | 2015-2019 | | | |
|-----------------------------------|-----------|-------------|---------|------------------------|------------|---------|-----------|------------|---------|--|
| | n | Mean (se) | P value | n | Mean (se) | P value | n | Mean (se) | P value | |
| Reviewer characteristics | | | | | | | | | | |
| Reviewer sex | | | | | | | | | | |
| Male | 284 | 65.5 (1.5) | .0022 | 188 | 64.6 (1.3) | .0001 | 311 | 70.1 (1.0) | .5548 | |
| Female | 132 | 71.5 (1.9) | | 317 | 69.6 (1.1) | | 631 | 71.3 (0.8) | | |
| Reviewer URM | | | | | | | | | | |
| Yes | 11 | 59.6 (6.4) | .1700 | 37 | 67.0 (2.6) | .7794 | 239 | 72.4 (1.1) | .0625 | |
| No | 260 | 68.63 (1.6) | | 437 | 67.7 (1.1) | | 693 | 70.6 (0.8) | | |
| Reviewer and applicant sex | | | | | | | | | | |
| Male reviewer | | | .3491* | | | | .4034* | | | |
| Male applicant | 136 | 65.1 (2.2) | | 87 | 65.9 (2.0) | | 147 | 67.6 (1.4) | | |
| Female applicant | 148 | 65.8 (2.1) | | 101 | 63.5 (1.8) | | 164 | 73.4 (1.3) | | |
| Female reviewer | | | | | | | | | | |
| Male applicant | 63 | 69.2 (2.8) | | 113 | 69.7 (1.8) | | 286 | 70.2 (1.2) | | |
| Female applicant | 69 | 73.6 (2.6) | | 204 | 69.5 (1.4) | | 345 | 72.3 (1.1) | | |
| URM reviewer and applicant | | | | | | | | | | |
| URM reviewer | | | .2082* | | | | .0104* | | | |
| URM applicant | 2 | 38.9 (14.6) | | 3 | 85.7 (8.5) | | 42 | 71.2 (2.6) | | |
| Non-URM applicant | 7 | 66.7 (7.8) | | 27 | 67.2 (2.9) | | 160 | 72.9 (1.3) | | |
| Non-URM reviewer | | | | | | | | | | |
| URM applicant | 25 | 68.5 (5.1) | | 47 | 66.1 (3.1) | | 82 | 66.4 (2.2) | | |
| Non-URM applicant | 144 | 74.1 (2.1) | | 313 | 70.6 (1.2) | | 539 | 71.2 (0.9) | | |

se, standard error.

*P value refers to the interaction term.

Furthermore, we could not include all applicants and reviews because of the challenges with records over time. Going forward, future applications and study section scores will be preserved to enable ongoing evaluation. Another limitation is that factors such as the track record of the mentor, the scientific momentum of the applicant, and the institutional environment contribute to the reviewers' scores. However, these are difficult constructs to quantify and were not captured during study section.

In conclusion, URM applicants received significantly lower scores than non-URM applicants. The impact of reviewer sex and URM status changed over time. Although male reviewers gave lower scores in the early periods, this effect did not persist in the latest period. Efforts are required to mitigate the impact of applicant URM status on application scores.

References

1. Todd RF 3rd, Gitlin SD, Burns LJ; Committee On Training Programs. Subspecialty training in hematology and oncology, 2003: results of a survey of training program directors conducted by the American Society of Hematology. *Blood*. 2004;103(12):4383-4388.
2. Sung L, Crowther M, Byrd J, Gitlin SD, Basso J, Burns L. Challenges in measuring benefit of clinical research training programs—the ASH Clinical Research Training Institute example. *J Canc Educ*. 2014;30(4):754-758.
3. Burns LJ, Clayton CP, George JN, Mitchell BS, Gitlin SD. The effect of an intense mentoring program on junior investigators' preparation for a patient-oriented clinical research career. *Acad Med*. 2015;90(8):1061-1066.

4. Helman A, Bear A, Colwell R, eds. *Promising Practices for Addressing the Underrepresentation of Women in Science, Engineering, and Medicine*. Opening Doors; 2020.
5. Ioannidou E, Letra A, Shaddox LM, et al. Empowering women researchers in the new century: IADR's strategic direction. *Adv Dent Res*. 2019;30(3):69-77.
6. Schnierle J, Christian-Brathwaite N, Louisias M. Implicit bias: what every pediatrician should know about the effect of bias on health and future directions. *Curr Probl Pediatr Adolesc Health Care*. 2019;49(2):34-44.
7. National Institutes of Health. Populations Underrepresented in the Extramural Scientific Workforce. 2022. <https://diversity.nih.gov/about-us/population-underrepresented>
8. King AA, Vesely SK, Elwood J, Basso J, Carson K, Sung L. The American Society of Hematology Clinical Research Training Institute is associated with high retention in academic hematology. *Blood*. 2016;128(25):2881-2885.
9. King AA, Vesely SK, Vettese E, et al. Impact of gender and caregiving responsibilities on academic success in hematology. *Blood Adv*. 2020;4(4):755-761.
10. Onken J, Chang L, Kanwal F. Unconscious bias in peer review. *Clin Gastroenterol Hepatol*. 2021;19(3):419-420.
11. Ginther DK, Schaffer WT, Schnell J, et al. Race ethnicity, and NIH research awards. *Science*. 2011;333(6045):1015-1019.
12. McKenzie ND, Liu R, Chiu AV, et al. Exploring bias in scientific peer review: an ASCO initiative. *JCO Oncol Pract*. 2022;18(12):791-799.
13. Berger JG, Purton LE. Gender inequities in medical research funding is driving an exodus of women from Australian STEM academia. *Immunol Cell Biol*. 2022;100(9):674-678.
14. Fox CW, Paine CET. Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. *Ecol Evol*. 2019;9(6):3599-3619.