Publications of the Astronomical Society of the Pacific, 132:034503 (6pp), 2020 March https://doi.org/10.1088/1538-3873/ab6ce0

Dual-anonymization Yields Promising Results for Reducing Gender Bias: A Naturalistic Field Experiment of Applications for *Hubble Space Telescope* Time

Stefanie K. Johnson[1] and Jessica F. Kirk[2]
[1] Leeds School of Business, University of Colorado Boulder, Boulder, CO, USA; Stefanie.Johnson@colorado.edu
[2] Fogelman College of Business & Economics, University of Memphis, Memphis, TN, USA
Received 2019 October 20; accepted 2020 January 15; published 2020 February 18

Abstract

Using archival data, we examine the effects of the *Hubble Space Telescope* Time Allocation Committee (*HST* TAC)'s decision to adopt a dual- rather than single-anonymous review process. The change involved removing, to varying degrees, information about the Principal Investigator (PI) with the goal of reducing bias against women. Proposals led by female PIs were significantly more likely to be accepted in the five cycles following the changes compared to the 11 cycles using a single-anonymous review system. Taking a closer look at why these changes emerged, we examined data at the reviewer-level in the cycle immediately preceding the change compared to three of the cycles after the change. We found that male reviewers rated female PIs significantly worse than they rated male PIs before, but not after, dual-anonymization was adopted.

Key words: Sociology of Astronomy

Receiving access to funding and resources can be the determining factor in one's success as a researcher and academic. Many scholars argue that biases in grant review processes result in lower levels of funding for women compared to men (Pohlhaus et al. 2011; Shen 2013; Urry 2015; Guglielmi 2018; Mallapaty 2018), although others have failed to find gender differences in this regard (Ceci & Williams 2011; Forscher et al. 2019). However, even when differential funding rates between men and women are evident, the numerical difference does not mean that bias was at work. Alternatively, it is possible that there are gender differences in the quality of applications because women have less access to mentors, collaborators, and other resources in writing grant proposals (Ley & Hamilton 2008; Moss-Racusin et al. 2012; Knobloch-Westerwick et al. 2013; Larivière et al. 2013; Shen 2013; Caplar et al. 2017). The lack of clear evidence of discrimination is compounded by the fact that there are few interventions known to reduce gender bias (Galinsky et al. 2015; Breda & Hillion 2016; Tricco et al. 2017). Furthermore, there is an inherent risk that trying an intervention can elicit backlash from non-beneficiaries (Goldin & Rouse 2000). As a result, many funding organizations have not made substantive changes to reduce gender bias.

Yet, at least one study shows compelling evidence that small interventions can significantly reduce the impact of gender bias. In 2014, the Canadian Institutes of Health Research (CIHR) created two separate granting processes—one that focused primarily on the science and one that focused primarily on the scientist, including an assessment of their leadership, productivity, and the significance of their contributions (Witteman et al. 2019). An analysis of nearly 24,000 applications showed that women performed as well as men in the science-only review process but worse than men in the scientist review process. Importantly, the applicants self-selected into the different grant programs, meaning that there could have been differences in the types of researchers who applied each program. However, the findings are consistent with the theoretical argument that bias is more likely to occur when evaluating individuals (the scientist) rather than focusing on their work (the science) (Heilman & Caleo 2015).

The current study examines if (1) statistical bias exists between male and female PIs applying for funding and access to the *Hubble Space Telescope* (*HST*), (2) using a dual- rather than single-anonymous review system mitigates bias, and (3) there is any difference between male and female reviewers in the impact of bias and dual-anonymization. The expectations that male reviewers will exhibit more bias against female PIs than female reviewers is consistent with some past work, although there is also evidence to the contrary (Eagly et al. 1992; Beaman et al. 2012). The present findings show that (1) there is evidence of statistical gender bias in favor of men, (2)

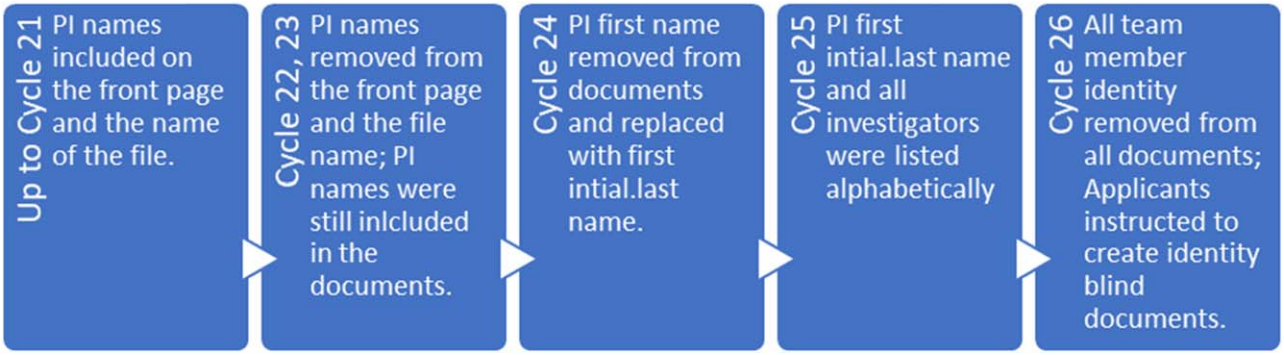

Figure 1. Stages of dual-anonymization at *HST* TAC—Boxes contain additional changes made in each cycle.

the gender bias was reduced following dual-anonymization, and (3) male reviewers rated female PIs significantly worse than they rated male PIs before but not after the adoption of dual-anonymization.

1. Methods

Each year, members of the astronomical community submit proposals for telescope time to the *HST* TAC. All proposals are sent to volunteer reviewers who rate their allotted proposals independently, then meet in small groups to decide on overall rankings and acceptance. For the last 16 cycles, *HST* TAC has recorded the relative success rate of male and female PIs. After finding evidence of statistical gender bias in the application process (Reid 2014), *HST* TAC changed their application procedures to reduce the salience of the PI's identity with the goal of reducing gender bias. Several attempts were made to improve the system, each with limited success, causing the *HST* TAC to continue to refine the application process.

In cycles 22 and 23, *HST* TAC removed PI names from the front page of the application document and file name but left the full names in the body of the application. In cycle 24, the PI's first name was replaced with a first initialin the body of the application, maintaining the other changes. In cycle 25, the names of the investigators (first initials and last names) were listed in alphabetical order so it would be difficult to identify which scientist was the PI. Finally, in cycle 26, all information about all investigators was removed completely and applicants were specifically instructed to write their proposals in a way that masked their identity (Figure 1). Our first set of analyses examine whether the success rate of male and female PIs differed in cycles 22–26 compared to cycles 15–21.

There were 15,545 applications across 16 yrs (or cycles) of data. Among those, 3533 proposals had a female PI. Across cycles, male PIs had an acceptance rate of 23% and female PIs had an acceptance rate of 19%, consistent with past research from *HST* TAC showing statistical gender bias (Reid 2014).

The second set of analyses focus specifically on cycle 21 (the cycle immediately preceding the changes related to dual-anonymization) and three of the cycles following dual-anonymization (cycles 24–26). In these cycles, *HST* TAC collected data at the reviewer level, allowing us to test whether the move to dual-anonymization had a greater impact on male or female reviewers. We use the ratings that the reviewers provided to *HST* TAC when they reviewed the applications on their own (not in groups). Unfortunately, these data were not available from *HST* for cycles 22 and 23. We test the effects of dual-anonymization, PI gender (0 = male, 1 = female), and reviewer gender (0 = male, 1 = female) on ratings of the applications. In cycle 21, there were 806 male PIs and 288 female PIs. In the other cycles combined there were 2054 male PIs (cycle 24 = 826, cycle 25 = 877, cycle 26 = 351) and 737 female PIs (cycle 24 = 270, cycle 25 = 329, cycle 26 = 138). The resulting data set included 3884 applications with an average of 6 reviewers per applicant, resulting in 25,069 rows of data. To control for variation in ratings of the reviewers (i.e., some reviewers may give generally higher ratings than others), we *Z*-scored each reviewer's ratings across the applications they rated. This means that we are examining reviewers' relative ratings of applications. The ratings are given on a 1–5 scale, with 1 being the best. Therefore, higher ratings indicate worse ratings.

2. Results

All analyses use two-tailed significance tests. Data were analyzed using the Mixed Model function in SPSS (IBM Corp 2018). This analysis, commonly used in the social sciences (Klein & Kozlowski 2000), accounts for both random effects and fixed effects in predicting a continuous outcome variable that approximates a normal distribution, like the data reported here. Fixed effects are effects that affect the entire population of data and random effects affect only subsets of the data. Random effects often become important when the data is multilevel, or "nested" in groups. For example, data regarding the wellbeing of school children from an elementary school may be nested in grade and classroom. The resulting multilevel data has both fixed and random effects. The random effects are

any effects affecting a subset of the data, such as grade (second graders are generally worse off than first graders) or classroom (one teacher generally has happier children than another). The fixed effects are effects that can affect the entire population of schoolchildren such as gender, race, or socioeconomic status. Our data is multilevel, nested in cycle or application, so we must account for the random effects impacting only subsets of the data in order to better estimate the fixed effects of interest. The analysis uses a restricted maximum likelihood estimate to fit the model. Maximum likelihood estimates produce a statistical model that makes the observed data most probable. A restricted maximum likelihood estimate uses a likelihood function that negates the effect of nuisance parameters.

Our first analysis examined the effect of PI gender (fixed effect) and anonymization (fixed effect) on the average success ratio of applicants, including a random effect for cycle (to account for any effects impacting only certain cycles) and including the overall success rate of applicants in each cycle as a covariate because some cycles had higher success rates than others. We used a multilevel file with two rows for each cycle (one for men, one for women) and each cycle coded as 0 = single-anonymized, 1 = dual-anonymized. In the analysis, we examined both the main effect of gender and cycle and the interaction between them.

The main effect is the estimated fixed effect of the independent variable (e.g., gender, cycle) on the dependent variable (e.g., success ratio), across the other independent variables. For example, the main effect of PI gender is the effect of PI gender on success ratio, averaged across all cycles. A significant interaction means that the effect of one independent variable changes the effect of another independent variable. The nature of an interaction is better understood by looking at the simple effects of each variable. The simple effects are tests of the effect of one independent variable at specific levels of the other independent variable. For example, looking at the effect of gender in a specific cycle or the effect of cycle for a specific gender.

We first examined the main effects of adopting a dual-anonymized approach (cycles 11–21 = 0, cycles 22–26 = 1) and PI gender (0 = men, 1 = women). As shown in Figure 2 and Table 1, there was no main effect for changing to a dual-anonymized system on overall acceptance rates, but there was a significant effect of PI gender ($B = -0.04$, df = 28, SE = 0.01, $p < 0.01$, 95% CI [−0.054, −0.029]), such that women experienced lower rates of success than men across all 16 cycles. Note, the term B is used to denote the estimate of the effect given in the analysis. We then used the estimated marginal means and standard errors from the mixed model analysis to calculate the estimated effect size, which was large (Cohen's $d = 2.63$). Additionally, df notes the degrees of freedom, SE is used for standard error, and both the p value and the 95% confidence interval (CI) are used to note significance.

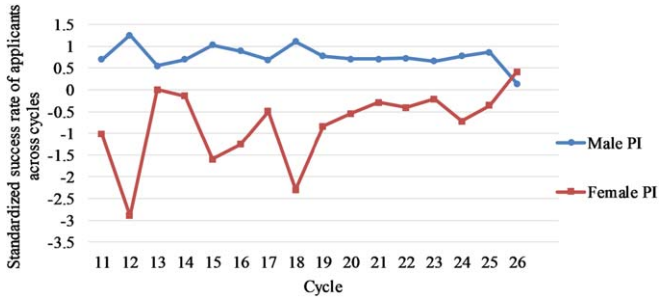

Figure 2. Plot of the standardized residuals of the success rate (percent funded divided by percent applied by gender) over the last 16 application cycles at *HST* TAC controlling for overall percent accepted at each cycle. The blue line represents the acceptance rate for male PIs and the red line represents the acceptance rate for female PIs. *HST* TAC began making changes to the application process in Cycle 22, although full dual-anonymization was not adopted until Cycle 26.

We then tested and found a significant interaction between PI gender and the adoption of dual-anonymization ($B = 0.03$, df = 27, SE = 0.01, $p < 0.05$, 95% CI [0.001, 0.051]). The full results are shown in Table 2. Looking at the simple effect of dual-anonymization (the effect of dual-anonymization for each gender) showed no effect for male PIs ($B = -0.01$, SE = 0.01, $p > 0.05$, 95% CI [−0.023, 0.013]), but a significant increase in the success rate of female PIs following the change to dual-anonymization ($B = 0.02$, SE = 0.01, $p < 0.05$, 95% CI [0.003, 0.039]). We used the estimated marginal means and standard errors from the mixed model analysis to calculate the effect size, which was large (Cohen's $d = 0.84$). Looking at the simple effect of gender, we see that male PIs had higher success rates than female PIs in cycles before (cycles 11–21; $B = -0.05$, SE = 0.01, $p < 0.01$, 95% CI [−0.064, −0.036], Cohen's $d = 3.02$) and after dual-anonymization was adopted (cycles 22–26; $B = -0.02$, SE = 0.01, $p < 0.05$, 95% CI [−0.045, −0.003], Cohen's $d = 1.53$), but this effect was weaker in the dual- rather than single-anonymized cycles.

Based on the evidence that (1) a statistical bias existed between the acceptance rates of male and female PIs and (2) dual-anonymization interventions in cycles 22–26 reduced this difference by significantly increasing female PIs' success rates, we dove further into the data to test whether there is any difference between male or female reviewers in the impact of dual-anonymization. We investigate the effects of adopting dual-anonymization, PI gender (0 = male, 1 = female), and reviewer gender (0 = male, 1 = female) on individual ratings of applications.

Changes in gender bias over time for male and female reviewers. As with the first analysis, we analyzed the data using the Mixed Model function in SPSS (IBM Corp 2018) to fit a linear mixed model with both fixed and random effects. The

Table 1
Main Effects of Blinding Intervention and PI Sex

DV = Success Ratio	*B*	SE	*t*	*p*-value	95% CI
Total Ratio	**00.982**	**0.066**	**14.812**	**0.000**	**0.846, 1.117**
Blind	0.008	0.006	1.235	0.227	−0.005, 0.022
PI Sex	**−0.042**	**0.006**	**−6.868**	**0.000**	**−0.054, −0.029**

Note. Bold numbers indicate a significant estimate of the effect (*B*), meaning the *p*-value is less than 0.05 and the 95% confidence interval (CI) does not cross zero. SE indicates the standard error and *t* is the value of the *t*-test significance test.

Table 2
Interaction Effect of Blinding Intervention and PI Sex

DV = Success Ratio	*B*	SE	*t*	Sig	95% CI
Total Ratio	**0.982**	**0.063**	**15.689**	**0.000**	**0.854, 1.111**
Blind	−0.005	0.009	−0.536	0.597	−0.023, 0.013
PI Sex	**−0.050**	**0.007**	**−7.206**	**0.000**	**−0.064, −0.036**
Blind X PI Sex	**0.026**	**0.012**	**2.101**	**0.045**	**0.001, 0.051**

Note. Bold numbers indicate a significant estimate of the effect (*B*), meaning the *p*-value is less than 0.05 and the 95% confidence interval (CI) does not cross zero. SE indicates the standard error and *t* is the value of the *t*-test significance test.

fixed effects, or the effects of interest impacting the entire population of data, were PI gender, reviewer gender, and cycle. The analysis was done at the reviewer level in order to observe reviewer effects. As such, the data includes multiple lines for each application and a random effect for application. Including this accounts for any differences between applications by modeling any effects that impact only one application. Since reviewers also reviewed multiple applications, the data has multiple lines for each reviewer. To account for differences between reviewers (one reviewer just generally scores higher), we *z*-scored the ratings of each individual rater (so a given score is relative to all other scores that reviewer gave). This is a simpler alternative to including reviewer as an additional random effect. In this case, we accounted for the one effect that mattered, how generally high or low a reviewer rates.

We removed outliers (deleting them from the data set) in the ratings using the interquartile range based on the full sample of ratings. The results did not change after removing the outliers. The covariates of PhD completion years of both the PI and the reviewer were included in the analyses. In cases where PhD year was missing, we inserted the grand mean of the PhD year. Given that we had multiple cycles of data, we created dummy variables for each cycle. Dummy variables are numerical variables with values of 0 or 1 that represent the presence or absence of a category. For example, the dummy variable for cycle 21 would be 1 for cycle 21 and 0 for all other cycles. We used the base approach for analysis with dummy variables (Yip & Tsang 2007) to compare cycle 21 to the other three cycles. The independent variables were reviewer gender, PI gender, and cycle 21 (the single-anonymized cycle).

The analysis included main effects of all independent variables. We created two-way interactions for reviewer gender by PI gender, reviewer gender by cycle 21, and PI gender by cycle 21. We also created a three-way interaction between reviewer gender, PI gender, and cycle. Three-way interactions test whether one variable changes the interaction between two other variables. As with two-way interactions, these are best understood by looking at the simple effects of a variable at levels of the other two variables. As with the first analysis, we analyzed the data using the Mixed Model function in SPSS (IBM Corp 2018) to fit a linear mixed model with both fixed and random effects. The fixed effects, or the effects of interest impacting the entire population of data, were PI gender, reviewer gender, and cycle. Application was set as a random effect, or an effect that impacts subsets of the population. This is estimated because the analysis is done at the reviewer level, meaning there are multiple observations per application.

We employed the analysis described above and found that the three-way interaction was statistically significant ($B = -0.11$, df = 23,261, SE = 0.06, $p < 0.05$, 95% CI [−0.222, −0.001]). Figure 3 and Table 3 shows the results. The simple effects show that in cycle 21, male reviewers rated female PIs significantly worse than they rated male PIs ($B = 0.09$, SE = 0.04, $p < 0.05$, 95% CI [0.017, 0.160]). We used the estimated marginal means and standard errors from the mixed model analysis to calculate the effect size, which was small (Cohen's $d = 0.01$). However, they rated female and male PIs equally well in the dual-anonymized cycles. This indicates that adopting dual-anonymization successfully eliminated bias exhibited by male reviewers toward female PIs.

Table 3
Interaction Effect of Cycle 21, Rater Sex, and PI Sex

DV = Preliminary Ratings	*B*	SE	*t*	Sig	95% CI
PI PhD	−0.003	0.001	−4.311	0.000	−0.004, −0.002
Rater PhD	**0.000**	**0.001**	**0.484**	**0.628**	**−0.001, 0.001**
Cycle 24	−0.017	0.029	−0.591	0.554	−0.074, 0.040
Cycle 25	0.000	0.029	−0.010	0.992	−0.056, 0.056
Cycle 21	−0.042	0.033	−1.284	0.199	−0.106, 0.022
Rater Sex	0.004	0.017	0.241	0.809	−0.029, 0.037
PI Sex	0.008	0.027	0.280	0.779	−0.045, 0.061
Rater Sex X PI Sex	0.016	0.033	0.496	0.620	−0.048, 0.081
Cycle 21 X Rater Sex	0.024	0.029	0.821	0.412	−0.033, 0.080
Cycle 21 X PI Sex	0.081	0.045	1.788	0.074	−0.008, 0.169
Cycle 21 X Rater Sex X PI Sex	**−0.111**	**0.056**	**−1.977**	**0.048**	**−0.222, −0.001**

Note. Bold numbers indicate a significant estimate of the effect (*B*), meaning the *p*-value is less than 0.05 and the 95% confidence interval (CI) does not cross zero. SE indicates the standard error and *t* is the value of the *t*-test significance test.

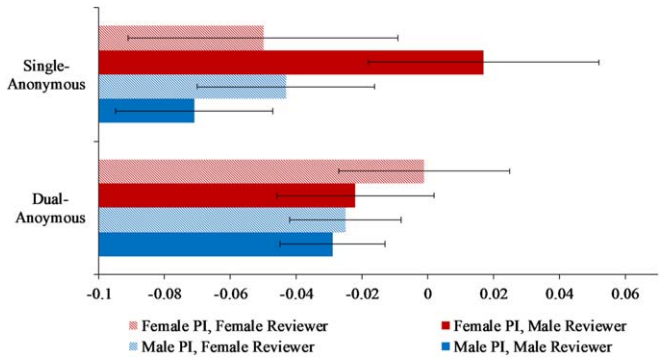

Figure 3. Three-way interaction of reviewer gender, PI gender, and dual-anonymization predicting ratings at the reviewer level. Male reviewers rated female PIs significantly worse than they rated male PIs in Cycle 21 but not in the dual-anonymized cycles. Higher ratings are equivalent to worse evaluations. Error bars indicate the confidence interval of the estimated means.

3. Discussion

There is mounting evidence of gender bias in the evaluation of women in science (Tricco et al. 2017). Although many fields have been slow to change, the astronomical community has been on the forefront of acknowledging gender bias and identifying ways to reduce bias (Reid 2014; Lonsdale et al. 2016; Patat 2016). The study reported herein represents a massive shift in the largest space telescope review process. Using a sample of 15,545 applicants over 16 review cycles, we show that female PIs were less likely than male PIs to receive access to telescope time when the review process was single- rather than dual-anonymized. Moreover, the analysis of 4 cycles of data at the reviewer level showed that male reviewers rated female PIs worse than male PIs before but not after dual-anonymization was adopted. Although using a dual-anonymized system (often called blinding) is becoming more common in industry settings, previous research investigating the impacts of dual-anonymization was limited.

Our findings support the case for dual-anonymization in the *HST* reviews, but generalize to grant proposals, conference presentations, publications, and employment. While many programs have been designed to help support women and minorities, they present two problems. First, very few have proven to be effective because unconscious gender bias is so automatic and difficult to overcome (Galinsky et al. 2015; Breda & Hillion 2016). As such, common interventions such as unconscious bias training do not seem to work over time. Instead, structural changes—such as increasing transparency—are more effective than trying to change individuals' reactions (Tricco et al. 2017). Second, many interventions cause backlash against women because women are perceived as receiving extra advantages or preferential affirmative action (Goldin & Rouse 2000). Using a dual-anonymization approach overcomes both of these obstacles (1) because dual-anonymization eliminates the possibility for bias to occur, rather than trying to overcome it, and (2) because it is difficult to argue that removing names from proposals is giving an unfair advantage to anyone.

There are several strengths to this research including the longitudinal design, a large sample size, and the use of a quasi-field experiment in a national agency. To change an entire selection process at a major national agency is not an easy task, as processes and procedures are often entrenched in bureaucracy. The findings reported here have important practical implications for all areas of science and academia. Insofar as we admit that bias against women exists, we are all responsible for intervening to stop it. Biases that impede the success of women in science limit the potential for innovation, remove important role models that diminish the pipeline of women in science, and create an impediment to social justice. With clear evidence that dual-anonymization mitigates bias of male

reviewers toward female principal investigators, there is little question that dual-anonymization should be widely considered in proposals for grants, publications, and even employment.

When there are differences between men and women in success, it is always difficult to unequivocally state that such differences are due to bias. This study provides very strong evidence that bias has, in fact, impacted the success of female scientists, at least in the context of *HST*. Dual-anonymization creates the most equitable outcome for all scientists. Further, unlike other interventions that may create the perception that achievement was not due to merit, dual-anonymization makes it possible for women to be treated equally.

The data presented here were received directly from the Space Telescope Science Institute (STScI). The relevant contact, Neill Reid, can be reached at inr@stsci.edu. The authors would like to thank the STScI for their full support and assistance on this research project.

Data Availability Statement

All data were received from *HST* TAC and will readily be shared by the authors upon request.

References

Beaman, L., Duflo, E., Pande, R., & Topalova, P. 2012, Sci, 335, 582
Breda, T., & Hillion, M. 2016, Sci, 353, 474
Caplar, N., Tacchella, S., & Birrer, S. 2017, NatAs, 1, 0141
Ceci, S. J., & Williams, W. M. 2011, PNAS, 108, 3157
Eagly, A. H., Makhijani, M. G., & Klonsky, B. G. 1992, Psychological Bulletin, 111, 3
Forscher, P. S., Cox, W. T., Brauer, M., & Devine, P. G. 2019, Nature Human Behavior, 3, 257
Galinsky, A. D., Todd, A. R., Homan, A. C., et al. 2015, Perspectives on Psychological Science, 10, 742
Goldin, C., & Rouse, C. 2000, American Economic Review, 90, 715
Guglielmi, G. 2018, Natur, 554, 14
Heilman, M. E., & Caleo, S. 2015, The Oxford Handbook of Workplace Discrimination (Oxford: Oxford Univ. Press)
IBM Corp 2018, IBM SPSS Statistics for Windows, Version 26.0 (Armonk, NY: IBM Corp.)
Klein, K. J., & Kozlowski, S. W. 2000, Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions (San Francisco, CA: Jossey-Bass)
Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. 2013, Science Communication, 35, 603
Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. 2013, Natur, 504, 211
Ley, T. J., & Hamilton, B. H. 2008, Sci, 322, 1472
Lonsdale, C. J., Schwab, F. R., & Hunt, G. 2016, arXiv:1611.04795
Mallapaty, S. 2018, Natur, 561, S9
Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. 2012, PNAS, 109, 16474
Patat, F. 2016, Msngr, 165, 2
Pohlhaus, J. R., Jiang, H., Wagner, R. N., Schaffer, W. T., & Pinn, V. W. 2011, Academic Medicine, 86, 759
Reid, I. N. 2014, PASP, 126, 923
Shen, H. 2013, Natur, 495, 22
Tricco, A. C., Thomas, S. M., Antony, J., et al. 2017, PLoSO, 12, e0169718
Urry, M. 2015, Natur, 528, 471
Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. 2019, The Lancet, 393, 531
Yip, P. S., & Tsang, E. W. 2007, Strategic Organization, 5, 13