OBSERVATIONS AND ANALYSIS OF DUAL-ANONYMIZATION AT HUBBLE SPACE TELESCOPE

Dr. Jessica F Kirk
University of Memphis
Fogelman College of Business & Economics

MY INVOLVEMENT WITH HUBBLE TAC

Analysis

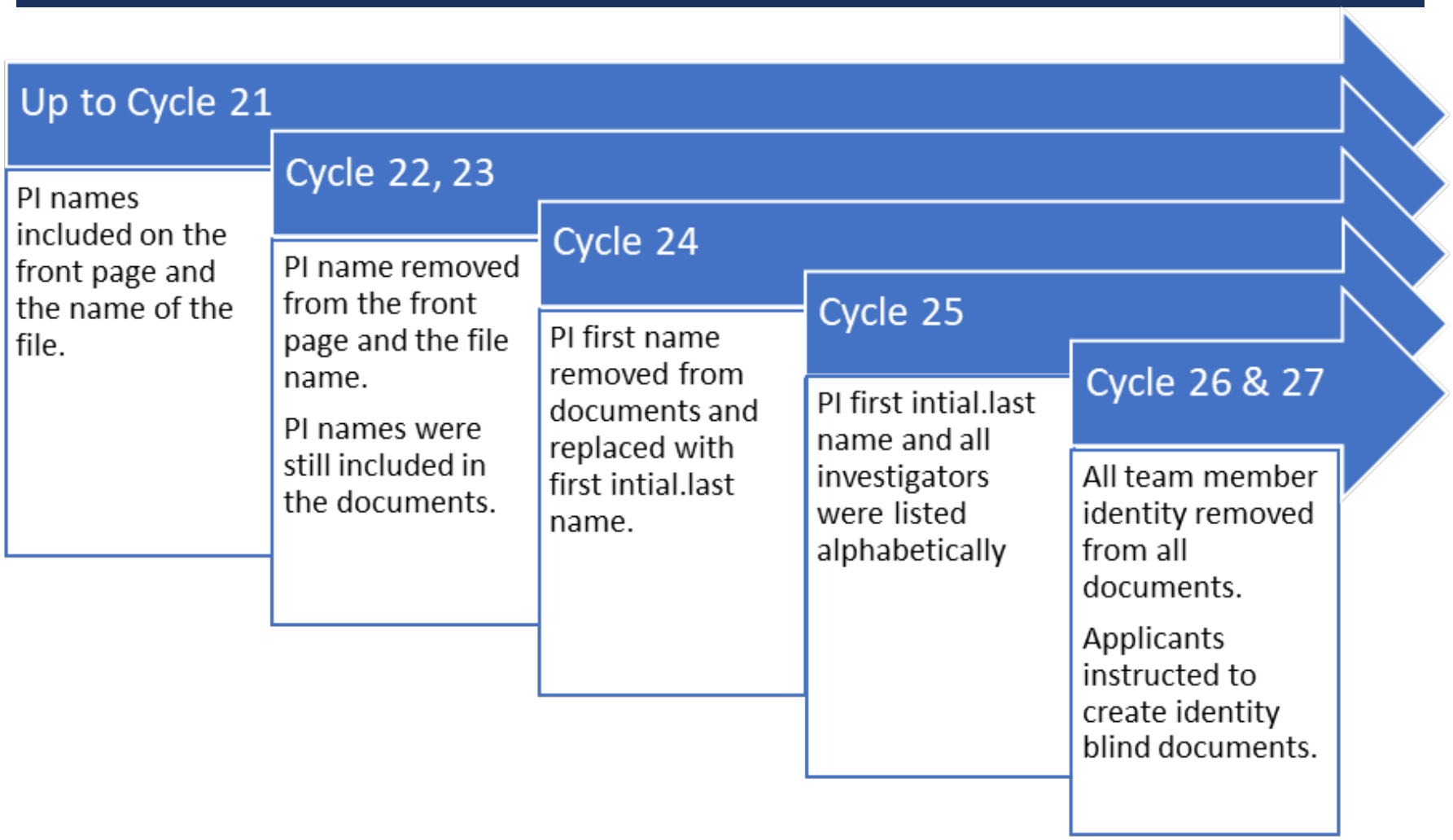
- Acceptance level data from Cycles 11-27
- Rater-level data from Cycles 21, 24, 25, & 26

Johnson, S. K., & Kirk, J. F. (2020). Dual-anonymization Yields Promising Results for Reducing Gender Bias: A Naturalistic Field Experiment of Applications for Hubble Space Telescope Time. *Publications of the Astronomical Society of the Pacific*, *132*(1009), 034503.

Observations

- Cycle 25 TAC in 2017
- Cycle 26 TAC in 2018
- Cycle 27 TAC in 2019

DUAL-ANONYMOUS INTERVENTIONS



OVERALL QUANTITATIVE ANALYSIS

Does a statistical bias exists between male and female PIs and does a dual-anonymous intervention effectively mitigate bias?

- Looking at acceptance rates across cycles 11 to 27
  - Men had an acceptance rate of 23%
  - Women had an acceptance rate of 19%

- Is this statistically significant?

METHOD

- Multilevel data with acceptance rates for men and women for each cycle from 11-27

- Mixed model with cycle as a random intercept
- Maximum likelihood estimation
- Controlled for overall acceptance rate in each cycle

- PI sex: 0 = men; 1 = women
- Intervention: 0=11-21; 1 = 22-27

RESULTS

- Main effect of PI Sex = -.04, SE = .01 $p < .01$, 95% CI [-.052, -.029]
- PI Sex X Intervention = .03, SE = .01 $p < .05$, 95% CI [.003, .049]

- Effect of Intervention – differences from 11-21 to 22-27
  - Men = -.00, SE = .01 $p > .05$, 95% CI [-.021, .012]
  - **Women = .02, SE = .01 $p < .05$, 95% CI [.005, .038]**
- Effect of PI sex – differences between men and women
  - **Cycles 11-21; B = -.05, SE = .01 $p < .01$, 95% CI [-.064, -.036]**
  - **Cycles 22-27; B = -.02, SE = .01 $p > .05$, 95% CI [-.042, -.005]**

APPLICANT SUCCESS RATES OVER TIME

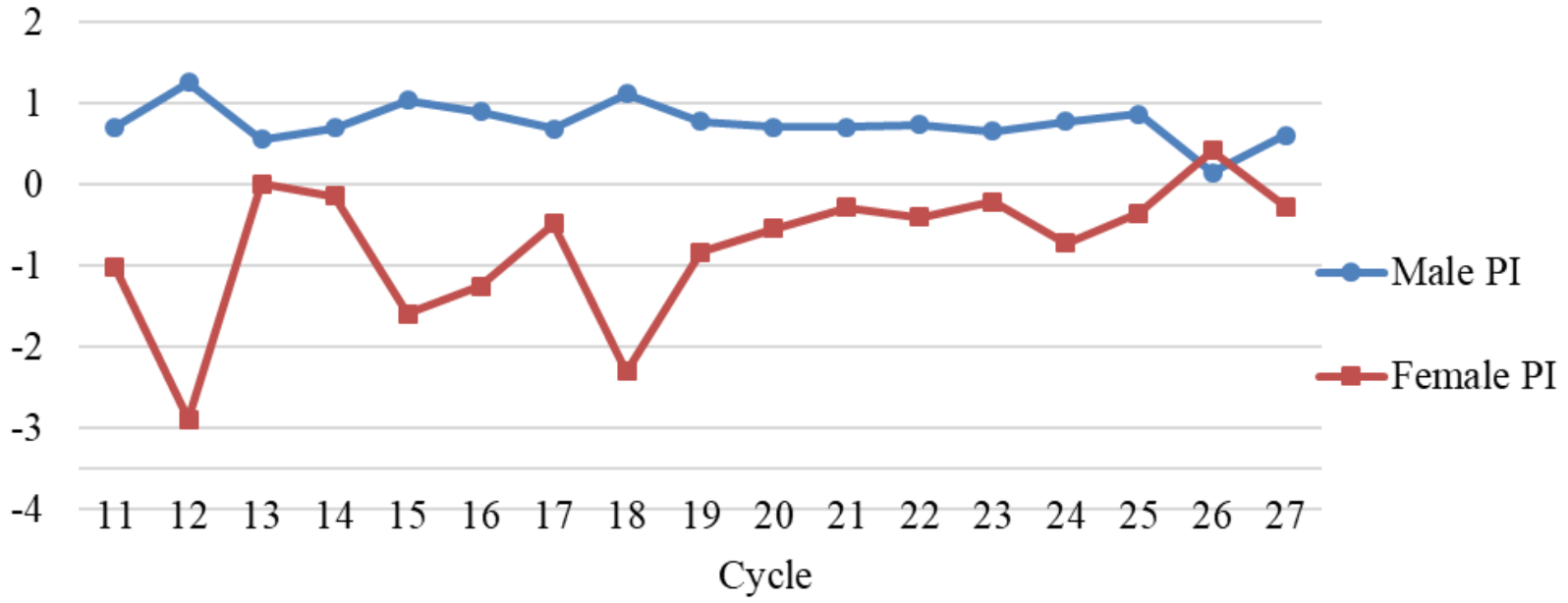


Figure 1: Residuals of the success rate (percent funded divided by percent applied by gender) over the last 16 application cycles at HSTAC controlling for overall percent accepted at each cycle. The blue line represents the acceptance rate for male PIs and the red line represents the acceptance rate for female PIs. The dashed vertical line indicates when HSTTAC started the blinding intervention.

DETAILED RATER LEVEL ANALYSIS

Are there any differences between male or female raters in the impact of the blinding intervention?

- Data at the rater level
  - 3,884 applications w/ average of 6 reviewers
  - 25,069 rows of data
- Cycle 21, Cycles 24-26

APPLICANT SUCCESS RATES OVER TIME

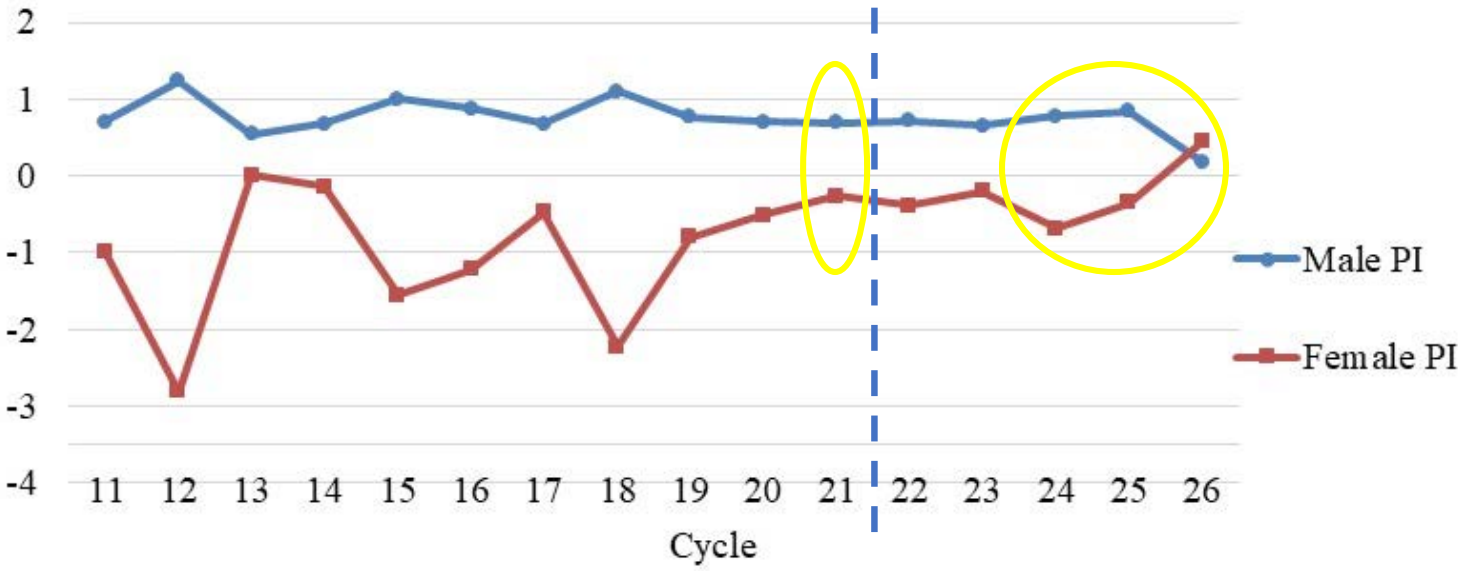


Figure 1: Residuals of the success rate (percent funded divided by percent applied by gender) over the last 16 application cycles at HSTAC controlling for overall percent accepted at each cycle. The blue line represents the acceptance rate for male PIs and the red line represents the acceptance rate for female PIs. The dashed vertical line indicates when HSTTAC started the blinding intervention.

METHOD

- Multilevel data with individual ratings for all applications from cycles 21, 24-26

- Mixed model with applicant as a random intercept
- Maximum likelihood estimation
- Controlled for overall acceptance rate in each cycle

- PI sex: 0 = men; 1 = women
- Rater sex: 0 = men; 1 = women
- Intervention: 0 = 21; 1 = 24-26

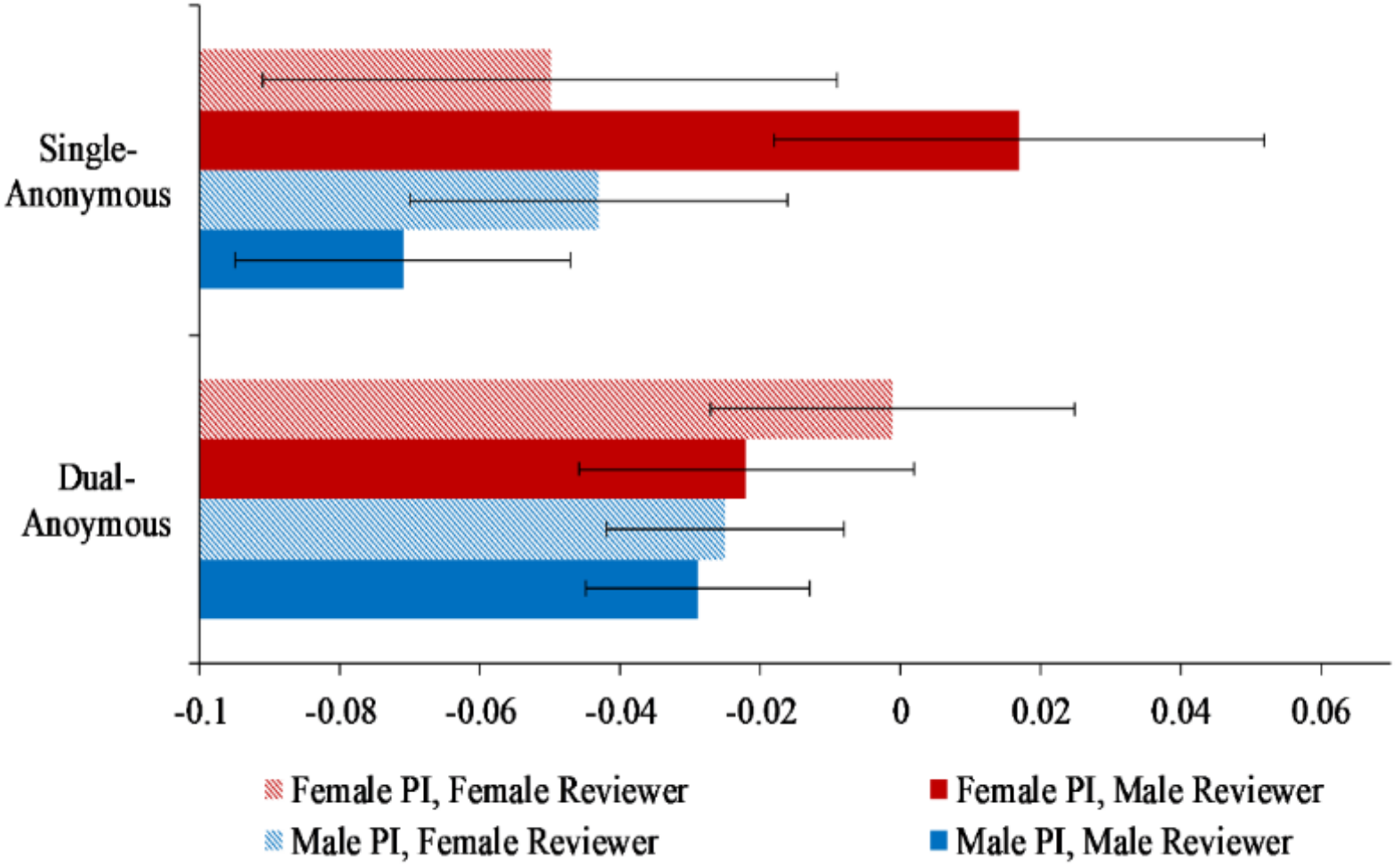
EFFECT ON PRELIMINARY RATINGS

- Effect on preliminary ratings (z-score)
- Comparing cycle 21 to cycles 24-26
- Control for PI and rater PhD completion year

- Three way interaction: PI sex X rater sex X cycle
  - B = -.11, SE = .06 $p < .05$, 95% CI [-.222, -.001]

- In cycle 21, male raters rated female PIs lower than male PIs
  - B = .09, SE = .04 $p < .05$, 95% CI [.017, .160]

EFFECT ON PRELIMINARY RATINGS



EFFECT ON FINAL RATINGS

- What happened in the panel meetings?
  - Effect on final ratings, controlling for preliminary (z-score)
- Effect partitioned out across cycles and raters
- Controlling for PI and rater PhD completion year

- Effect of PI sex for female raters in cycle 25
  - B = .11, SE = .05 $p < .05$, 95% CI [.004, .216]

POSSIBLE OVERCORRECTION

- Could this be a backlash or overcorrection of female raters?
- In cycle 25, Stefanie and I observed only 13 of the 16 panels. If the effect for female raters is an overcorrection, we would see this in the observed panels
- Effect of PI sex for female raters in observed panels
  - B = .17, SE = .07 $p < .05$, 95% CI [.036 .300]

QUALITATIVE OBSERVATIONS

Panel observations

- Cycle 25 TAC in 2017
- Cycle 26 TAC in 2018
- Cycle 27 TAC in 2019

CYCLE 25 OBSERVATIONS

- Observed 13 of the 16 panels in cycle 25
- Conducted qualitative analysis where we coded discussions about applications for specific elements
- Observed that approximately 50% of the conversations included some reference to the team, PI, lab, etc
- This may indicate a break down in the bias reducing effects of anonymizing seen in the preliminary ratings.

CYCLE 26 AND 27 OBSERVATIONS

- Observed all panels in cycle 26 and several in cycle 27
- No coded references to specific team, PI, lab, etc
- General positive attitude towards dual-anonymous
- Any issues were more process related

CONCLUSIONS

- Limited results on intervention indicate good news
- "Partial" implementations of dual-anonymous intervention may cause problems
- Overall, the full dual-anonymous cycles 26 and 27 ran smoothly!