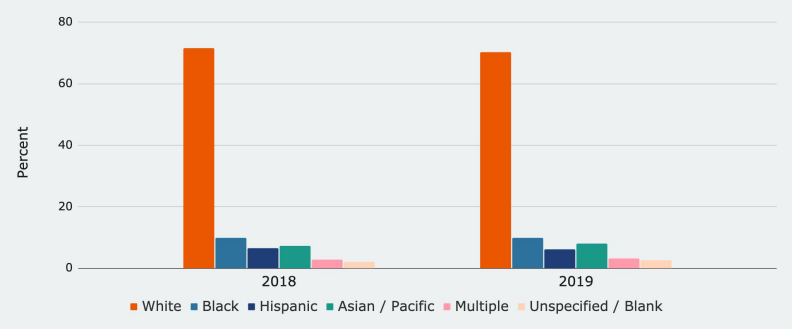
Is There Hidden Bias In Performance Evaluations?

The NewsGuild of New York, The New York Times

What Did We Look At?

- We analyzed how ratings from Guild members' performance reviews corresponded to racial and ethnic self-identification categories, gender, age and department.
- Widespread performance reviews began only in 2018; we examined 2018 and 2019 results.
- We got data on reviews from Guild members only. There were 986 in 2018 and 1,010 in 2019.
- Each year, the vast majority of members — about 86 percent — were in Newsroom or Opinion.
- About 70 percent of total members each year were white, and racial makeup was stable.



The Ratings System Over Time

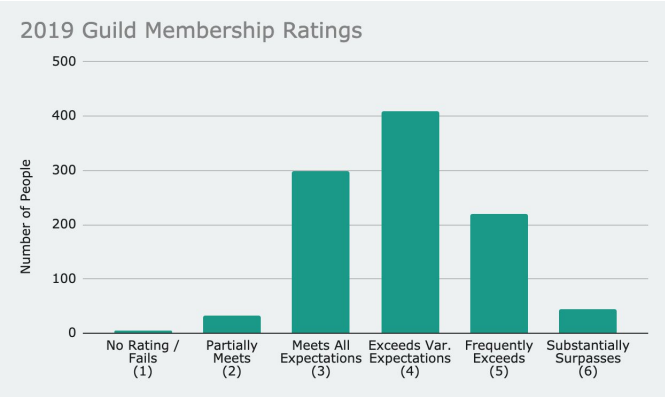
- The ratings system changed slightly from its first year to its second.
  - In 2018, there were five categories:

    Not Meeting Expectations, Building Performance, Solid Performance,
    Strong performance and Leading Performance.
  - In 2019, there were six categories:

    No Rating OR Failing Expectations, Partially Meets Expectations, Meets All Expectations,
    Exceeds Various Expectations, Frequently Exceeds Expectations and Substantially Surpasses Expectations.
  - In 2018, more than 40 employees received no rating. In 2019, the company emphasized the importance of reviews, and almost all employees received a rating.

The Ratings System Over Time

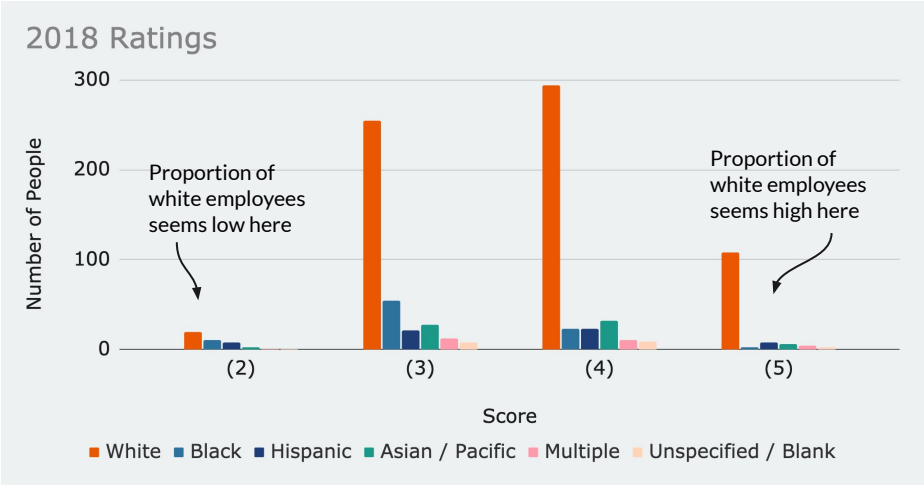
- In evaluating the ratings, we labeled the ratings using a scale from (1) to (5) in 2018 and (1) to (6) in 2019. For the primary analysis, we did not consider people who did not receive ratings. We also did not consider people who received what we labeled a (1), because there were only 2 in 2018 and 5 in 2019. These results are discussed in later notes.

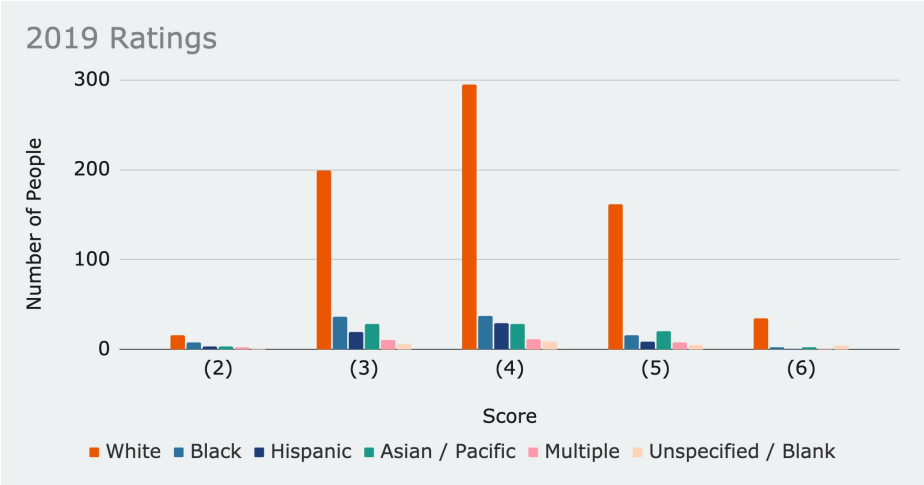


What Did We Find?

- There is a strong pattern of racial disparity in reviews. Employees of color were disproportionately likely to receive low ratings, while white employees were more likely to be rated highly.
- The discrepancies most clearly affect employees who identify as Black and Hispanic.
- Imbalances are obvious in 2019 as well as 2018 and occur in roughly the same pattern in both years.
- Although the data accessible to the Guild is limited, the results are strong enough that they warrant a clear call for additional investigation and transparency from management.

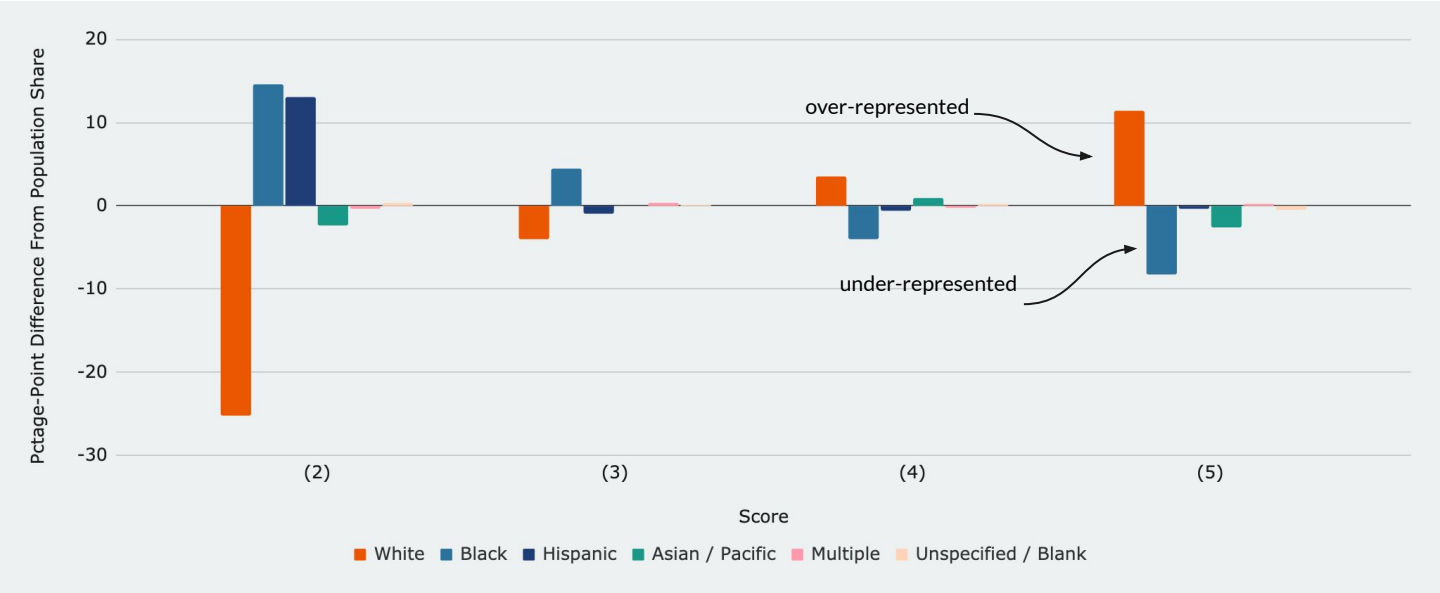
Raw Numbers Suggest Disparities But Are Overwhelmed by Population Differences





Compared with their share of the Guild population, workers of color are under-represented at high score levels and over-represented at lower ones.

How Each Racial Group Was Over-Represented Or Under-Represented at Each Score, 2018



How Each Racial Group Was Over-Represented Or Under-Represented at Each Score, 2019

To Put This Another Way

- In 2018, Black and Hispanic employees were 16.3 percent of membership but 43.9 percent of people scoring a (2). White members were 71.6 percent of the Guild but 83.1 percent of people with a (5), which was the top score.

- In 2019, Black and Hispanic employees were 15.9 percent of membership but 33.3 percent of people scoring a (2). White members were 70.3 percent of the Guild but 77.8 percent of people receiving a (6).

The Difference Is Statistically Significant

- In 2018, on average, Black employees scored 0.58 points less than white employees. The difference was significant at the 95 percent confidence level. Hispanic members scored 0.21 points less, on average, than white employees, and this too was statistically significant. Other nonwhite groups also scored lower, but the difference in those categories was not significant.

- In 2019, Black members scored 0.32 points less than white members, on average, a number that was again significant to 95 percent. Hispanic workers scored 0.23 points less; this was significant to the 90 percent level but not 95 percent. Most other groups also had lower scores than white members, on average, but the difference was not significant.

Notes: Analysis used OLS regression with categorical dummy variables for race. In 2018: R-square = 0.05; Significance F = 3.53E^-9 (statistically significant); 95% confidence range for each race's effect on score as compared to white employees as follows: Black: -0.75 to -0.42; Hisp: -0.41 to -0.01; Asian: -0.29 to 0.09; Multi: -0.39 to 0.20; Unknown: -0.46 to 0.21. In 2019: R-square = 0.02; Significance F = 0.006 (statistically significant); 95% confidence range: Black: -0.51 to -0.13; Hisp.: -0.46 to 0.01; Asian: -0.33 to 0.08; Multi: -0.45 to 0.19; Unknown: -0.16 to 0.56.

Notes: Employees Without Ratings

- In 2018, a significant number of employees did not receive a rating. Employees of color were slightly over-represented in this category. However, almost all employees received a rating in 2019.

- Two people, both Black employees, received a rating of "Not Meeting Expectations," a (1), in 2018. In 2019, this category was combined with those not receiving reviews at all, including some people leaving the company. This made analysis difficult. Five people received the 2019 version of the (1) rating: three white employees, one Hispanic employee and one whose race was not specified.

Notes: Gender and Age Analysis

- Gender analysis indicated that women were slightly less likely to receive the highest ratings. In 2018, they were 48 percent of the membership but only 40 percent of those in the top category. In 2019, women were 49 percent of Guild membership and 44 percent with the highest rating. Other ratings showed more balanced results*.

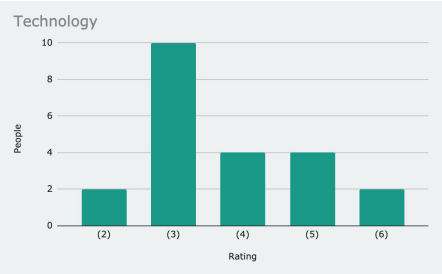
- We also analyzed discrepancies by age. Employees over 61 are less likely to receive the highest rating, but it is difficult to say this is compelling evidence of potential bias, considering that age, seniority, and job expectations may be logically correlated.

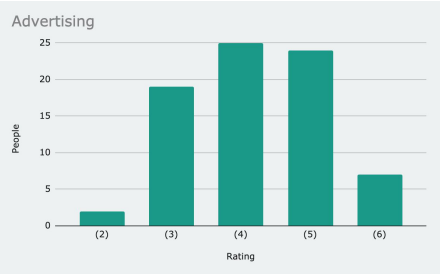
*Membership includes a small number of non-binary employees, but they did not receive the highest or lowest ratings.

Notes: Analysis by Department

- Some departments and desks may tend to provide skewed ratings. If members of certain races cluster in these sections, this may affect results. However, such inter-departmental explanations would *not* let the company off the hook. Rather, the immediate question should be: Why are departments with large numbers of employees of color rated more harshly?

  Technology and Advertising, for example, showed markedly different distributions in 2019. The large proportion of (3) ratings in Technology is particularly notable because that department has a relatively large percentage of employees who are Black.





Notes: Additional Caveats

- Because the data is restricted to ratings from Guild members, it is possible that it does not paint a full picture of the treatment of employees of different races and ethnicities when considering the company as a whole.

- The need to anonymize the data means that we can see only one attribute at a time. For example, we can tell how many Hispanic employees received a (4) and how many newsroom employees received a (4), but not how many Hispanic women in the newsroom received a (4). This means we cannot perform the sort of multivariate analysis that could help isolate potential biases.

- Guild members should be aware of the caveats above, but the pattern of racial disparities is so stark that it warrants additional attention by leadership in spite of any limitations in the analysis.