Biases and (Dis)agreement in Fellowship Selection Process: Insights and Strategies

Wichinpong "Park" Sinchaisri* & Titipat Achakulvisut†

February 27, 2018

1 Introduction

Review processes have been studied in multiple domains including employment interviews [1, 2], awarding of grants [3], and scientific peer review process [4, 5]. Studies have shown that they are prone to biases, for example, reviewers view applicants with ethnic names, speaking with an accent [2] or manuscripts written by authors who are further away from their co-authorship network less positively [5]. However, less is known about how the characteristics of the reviewers and the position (e.g., nature of the job, targeted audience of the journal) could affect the evaluation process. Moreover, when the evaluation outcome is a ranking rather than a binary accept/reject, will it be more prone to biases? The recent dataset of the fellowship selection process allows us to gain insights into these issues.

We analyze the dataset from Global Health Corps (GHC) fellowship review process, which includes more than 5,000 fellowship applications to GHC going through five rounds of reviews. This paper investigates the second and third rounds in which reviewers select 1,200 semifinalists from more than 4,000 applicants. We find significant characteristics that influence each applicant's ranking and how the citizenship of the reviewer and the position's skill requirement also play a role. We also study the source of disagreement among reviewers on their selection. We then propose strategies to reduce bias and inconsistency in evaluation: optimal reviewer assignment and machine learning-based selection.

2 Data and Methods

2.1 GFC Fellowship Review Process

The GHC fellowship application dataset consists of 5,778 applications to 139 fellowship positions. Minimum and maximum numbers of applicants to each fellowship position are 1 and 103 applicants, respectively. The distribution of number of applicants to fellowship positions is shown in Figure 2. There are 244 reviewers: 201 of them review for only one position, 34 review for two, 7 review for three, and 2 review for four. 4,220 applicants across 131 positions are reviewed by two reviewers in the second round. The distributions of citizenship and gender of applicants and reviewers are shown in Figure 1. The ratios of gender and citizenship between applicants and reviewers are similar. The majority of applicants are identified as Black (50.2%) followed by White (25.97%), Asian (10.81%), and Hispanic (3.46%). 67.04% of applicants are female and 32.49% are male. Figure 3 shows the number of applicants applying for a position requiring specific skills, color coded by gender. Other available characteristics include: citizenship of both applicants and reviewers, birth year, language proficiency, work experience in public health, and public health education.

Each reviewer assigns scores to various dimensions such as *commitment to social justice* and *innovation* for each applicant, sums them as a total score, and then ranks all the applicants to the same position (usually the top 10 are selected as semifinalists and additional 5 as alternates). We scale the total scores assigned by each reviewer to be within an interval between 0 and 1, where 0 represents the reviewer's lowest assigned score and 1 represents his/her highest. The scaled scores are denoted as *normalized scores*. Normalization allows us to

*Doctoral Candidate, Operations, Information, and Decisions, The Wharton School. swich@wharton.upenn.edu
†Doctoral Candidate, Department of Bioengineering, University of Pennsylvania. titipata@seas.upenn.edu

correct for reviewers' different levels of strictness/leniency and to compare them on the same scale. Further details of our pre-processing step can be found in Section A.1.

2.2 Empirical Methodology

We employ a variety of metrics to quantify how each applicant is viewed by reviewers: *rank* (the lower the better), *normalized score* (the larger the better), and two binary variables of *success rates*, one is whether the applicant was selected as a semifinalist or not, and the other is whether s/he was selected as at least an alternate (also the larger the better, for both). For each metric, we first run an ordinary least squares (OLS) model as a benchmark and then run a more appropriate model according to the observed distribution of such metric. We use negative binomial regression models to predict ranks (positive integers), beta regressions for normalized scores (values between 0 and 1), and Logit/Probit models for the success rates (binary 0 or 1).

For each model, we perform a stepwise model selection to select a set of characteristics with the best in-sample fit based on the Akaike Information Criterio (AIC)[1]. LASSO and other regularizations also suggest the same sets of features. To capture the heterogeneity among reviewers (e.g., their different standards), we analyze fixed effects models which allow for different intercepts among different reviewers. Lastly, to compare the level of (dis)agreement between the two reviewers, we run two different tests for each comparison of their characteristics: (i) t and Wilcoxon rank sum tests to compare the distributions of scores/ranks across each characteristic (e.g., same gender/citizenship or not), and (ii) regression models of the score/rank difference on whether the reviewers share any common characteristics.

3 Findings

3.1 Discrimination of Applicants' Demographics

Following [7], we first attempt to investigate whether race and gender biases exist in the fellowship review process. The naive way is to compute the ratio of applicants of a particular demographic who are selected to move forward. We observe that 60.31% of Whites, 51.27% of Blacks, 56.58% of Hispanics, and 54.79% of Asians are selected as at least an alternate by at least one reviewer. Compared to the racial breakdown presented in Section 2, it may appear that White applicants might have been reviewed much more positively than those of other races. However, such approach ignores the fact that *fellowship positions are not equally competitive*. Applicants applying to a more competitive position (i.e., there are more applicants) will have a smaller chance of being selected. After controlling for within-position competitiveness, we find that Black applicants actually have much better chance to be selected (53.94%), compared to White applicants (39.15%). Hispanics and Asians face even lower success rates of 36.61% and 25.04%, respectively. Similarly, to investigate gender bias, we find that 51.20% of male and 56.84% of female applicants move forward to Round 3. While it is true that females are more successful, the gap between males' and females' success rates is smaller when correcting for the position's competitiveness. Therefore, we conclude that *there is no systematic bias against Black or female applicants*. In fact, these applicants are reviewed more positively than others.

Consistent results from multiple regression analyses we described in Section 2.2 confirm our findings. Table 1 provides coefficient estimates from five different models: (1) negative binomial of ranks, (2) beta regression of normalized scores, (3) fixed effects model of normalized scores, (4) Logit model of being selected as a finalist, and (5) Logit model of being selected at least as an alternate. Across all models, male applicants are less favorable, while being eligible for the position (in terms of citizenship)[2], having work experience in public health, or having previously applied can boost their rankings, scores, and the likelihood of moving forward to the next round. We also observe that the effects of race are generally not significant, confirming our earlier conclusion that there is no (or very small if at all) racial discrimination in the reviewing process.

3.2 Roles of Reviewer's Demographics and Position's Characteristics

Biases can also arise when the applicants' demographics are similar to or different from the reviewer's and when the position requires a specific skill set. Here, we discuss four characteristics: citizenship, skill requirement, gen-

[1] AIC is a common measure for goodness of fit that favors smaller residual error but penalizes for including too many explanatory variables to avoid overfitting [6].

[2] Interestingly, 997 applicants applied for a job that they are not eligible for (because of their citizenship). However, 356 of them actually are selected to pass through to Round 3. All of them are rejected in Round 3.

der, and review preferences. Regression models discussed in the previous section include these characteristics as controls and the estimates are reported in Table 1.

Reviewers rank applicants of the same citizenship more positively than those of different citizenship. Table 1 shows a significant impact of mutual citizenship on each applicant's rank, probability of being selected as a semifinalist, and probability of being chosen as at least an alternate. In terms of normalized scores, while the estimates for the coefficient are not statistically significant, we observe a positive trend that applicants of the same citizenship as the reviewer score higher. Another interesting finding is that, *reviewers seem to be more generous in terms of assigning scores when reviewing applications to their home country, but the rankings and the probability of being selected are worse*, even when controlling for the number of applicants. This may suggest that reviewers have a higher standard when reviewing for their home country.

Skilled reviewers (possessing a skill required by the position they are reviewing for) are stricter than those without such skill. In other words, the rankings and the likelihood of being selected are worse than when reviewers do not have the required skill. Possessing the required skill may indicate that the reviewer has a deeper understanding or familiarity about the position's requirement. Thus, s/he may appear to have a higher standard when evaluating the applications. This effect is even stronger when they review applicants of the same citizenship; these applicants are 6-8% less likely to be selected to move forward.

The gender of the reviewer also matters, but in a different manner. Rather than favoring those with the same gender as their own, *male reviewers tend to be nicer than their female peers* as they are more likely to pass the applicants to the next round even though on average they assign lower normalized scores. We note that there are 2.6 times as many female reviewers as male reviewers and that male reviewers are also more likely to be assigned to positions popular among men.

Lastly, we investigate whether reviewers' preference in reading specific types of applications influences their evaluation. First, we observe that reviewers become slightly stricter when reviewing for the country where they are placed. However, the effect is only significant in the models of success rates. We further investigate by leveraging additional information about their review preference. Each reviewer expresses whether s/he prefers to read applications to her/his own organization or to review for others. 36.36% and 10.23% of the reviewers prefer to review for their organizations and other organizations, respectively. Although we observe only the position's placement country and the placement country of the reviewer, we can at least identify a subset of reviewers whose requests are fulfilled. We define "happy" reviewers as those who request to review for different organizations and are assigned to review for a different country. Those who wish to read for their institutions but are assigned to review for a different country are "disappointed". 24 reviewers are happy and 11 reviewers are disappointed. A Kruskal-Wallis test confirms that normalized scores are significantly different across happy, disappointed, and the rest of the reviewers. We find that *on average disappointed reviewers tend to give a higher normalized score, but with a larger variance.* In other words, disappointed reviewers tend to be less consistent and it might be difficult to interpret their evaluation.

3.3 (Dis)agreement among Reviewers

Here, we present only the analysis of positions that are reviewed by two reviewers. We compare within each pair of reviewers along five dimensions: gender, citizenship, placement country, skill set, and status (whether they're a fellow or alum). For each dimension, 50-60% of the positions recruit reviewers who share the same backgrounds, except that 80% of the positions recruit reviewers who share similar skill sets. Our metrics for (dis)agreement include mean and absolute difference of ranks and normalized scores, the number of applicants selected as semifinalists by both of them, and Spearman's rank correlation. Although the reviewers are instructed to review independently, we do find some similarities and quantify them as level of agreement.

Assigning the reviewers of the same gender to review for the same position does not have much significant impact on their evaluation. However, *reviewers of the same gender are significantly more likely to select the same applicants as semifinalists* ($W = 2466, p = 0.0217$[3]), which is also reflected by the larger rank correlation ($W = 2438, p = 0.0183$). However, on average, the average (raw or normalized) scores and ranks between each pair do not seem to be influenced by whether they are the same gender.

Reviewers of the same citizenship tend to agree more often. Wilcoxon rank sum tests confirm that these reviewers select a significantly larger number of same applicants as semifinalists ($W = 2430.5, p = 0.0149$) or at least alternates ($W = 2391.5, p = 0.0255$). The absolute difference between their assigned score is also smaller than

[3]We report only the results from Wilcoxon test as it is nonparametric and more robust. t test provides the same conclusions.

when they are from different countries. We also investigate the behaviors of reviewers who are placed into the same country but find no significant differences in their evaluation.

Finally, *sharing the same skill set or status does no have any significant impact on level of agreement between two reviewers*. We do, however, observe some qualitative patterns. Reviewers with the same skill set tend to disagree more often as their rank correlations and the number of applicants selected by both are smaller than when the review team with different skills. Similarly, reviewers of the same status (both alumni or both fellows) share fewer applicants who they both think should be moved forward to Round 3.

3.4 Reconciling Evaluations from Round 2 in Round 3

We divide applicants from Round 2 into two main groups based on whether they are recommended (as semifinalists or alternates) by two reviewers or just one. There are 831 and 1,231 applicants who are recommended by both reviewers and one reviewer in Round 2, respectively. Only 39 applicants selected in Round 3 without any suggestion from Round 2. If both reviewers recommend the applicant, there is a 82.07% chance that s/he will get selected again in Round 3. If only one reviewer recommends, the chance is decreased to only 39.2%.

Next, we further investigate how the reviewer in Round 3 decides on those with only one recommendation. We first compare the group of those selected by one reviewer with the group that Round 3 reviewer chooses. We find no significant difference in gender distribution between them (Kolmogorov-Smirnov test, $D = 0.04$, $p = 0.627$). In terms of citizenship, we observe that, among those with one recommender, Zambia, Uganda, and United States are the most represented with 44.55%, 42.94%, and 40.23%, respectively. Such distribution remains the same for the final selection (Mann-Whitney rank test, $D = 33, p = 0.106$). Since the majority of Round 3 reviewers are from the US (65.24%), we test whether reviewers seem to follow the recommendation from the reviewer of the same citizenship by estimating a Probit model of the final selection. The result confirms that there is no significant citizenship bias in this round. Similar analysis is performed to compare American and non-American reviewers, and again we find no significant differences in their selection. We also run similar tests as in Section 3.2 to confirm that Round 3 reviewers are not biased towards applicants with the same citizenship as them.

In addition, we include maximum and minimum normalized scores given by the two reviewers into our Probit models. We find that there is a significant correlation between the maximum of normalized scores and the likelihood of getting selected in Round 3 ($p < 0.001$, see Table 2). We obtain the same result when also including applicants who receive two recommendations ($p < 0.001$). Therefore, we conclude that *Round 3 reviewer generally selects applicants who are recommended by two previous reviewers first and then fills the remaining spots by ranking the normalized scores*.

4 Recommended Strategies

4.1 Optimal Reviewer Assignment

As we find that Round 3 reviewers tend to agree with the (mutual) choices of Round 2 reviewers and use normalized scores to determine the remaining spots, the goal is then to reduce biases and inconsistency of their evaluations. We first suggest that, while the reviewers can still keep the current scoring system, *normalized scores should be used when comparing the evaluations across different reviewers*.

In assigning a reviewer to a position, we should *avoid matching his/her citizenship with the applicants'*. This can be practically challenging because some opportunities are available to applicants of multiple citizenships (23.63% of positions are open to any applicants who are not a US citizen). In such case, we propose a *weighting scheme* to correct for potential mutual citizenship biases. For each reviewer i who reviews applicants from same and different citizenships, we look at other reviewers $j \neq i$ (referred to as *matched reviewers*) who share similar characteristics (e.g., gender, skill set, status). Let M and N be the sets of applicants who have matching and non-matching demographics with the reviewer's. Also, let s_i^j be applicant i's score assigned by reviewer j. We compute the means of normalized scores, one from matched reviewers who only reviewed applicants from their countries $\bar{s}_1 = \sum_{k \in M} s_k^j / |M|$, and another from those who only reviewed applicants of different cizitenships $\bar{s}_2 = \sum_{k \in N} s_k^j / |N|$, controlling for the applicants' demographics. Then, we can scale the biased scores of our focal reviewer by $\bar{s}_2 / \bar{s}_1$ for each group of applicants with the same demographics.

When building a review team, the organization needs to first identify whether it is more preferable to have a diverse set of semifinalists. Assuming that the goal is to have a larger consideration set for the later rounds

(e.g., fewer overlapping recommendations), it is better to *recruit reviewers with different citizenships and different genders*. One reviewer should be from the country of the position's placement (i.e., reading for a placement to home country). *At least one reviewer should possess the skill required by the job.* It is important that the review team should have a clear understanding of what the position requires and is able to evaluate the applicants accordingly. Lastly, *satisfying a reviewer's preference can help reduce the variance*. When reviewers do not get their preferred assignment, their evaluations could be harder to interpret due to increased variance.

4.2 Data-Driven Approach to Make Final Selection

As discussed in Section 3.4, we find that there is no significant bias in Round 3. We propose two systematic schemes to rank applicants in Round 3: *score ranking model* and *random forest-based ensemble*. We assess their performance by comparing the accuracy as if we replace Round 3 by either of them. Accuracy is computed as the number of semifinalists selected by both Round 3 reviewers and our algorithms and round 3 selections divided by the number of semifinalists selected by Round 3 reviewers.

Our first algorithm, *score ranking model*, selects applicants with two recommendations and then fills the remaining spots (if any) with those whose normalized scores are the highest. The second algorithm, *random forest-based (RF-based) ensemble*, employs a random forest model to predict the final selection based on mean and maximum of normalized scores, maximum total score, and two dummy variables: whether the applicant receives exactly one or two recommendations. To avoid over-fitting, we train our models on 30% of the dataset and construct 10 random forest models, each with 4 estimators and maximum depth of 5. Applicants are then ranked by their average success probability from the ensemble of all random forest models.

Global accuracy of the score ranking model is 73.84% whereas the RF-based ensemble gives a 77.32% accuracy. The accuracy distribution by placement country is illustrated in 4. To further quantify the performance of our proposed models, we compare it to two baselines, random selection and selection based on the maximum of average scores (i.e., applicants with highest average scores will be selected). Random selection and maximum of average scores-based selection yield 39.70% and 70.34% accuracy, respectively. Despite only trained with 30% of the data, RF-based ensemble outperforms other methods with almost 8 of 10 candidates are correctly selected. Therefore, we suggest that *RF-based ensemble can be used in lieu of human reviewers in Round 3, shortening the review process time and reducing labor cost while making the same final selection*.

5 Conclusions and Directions for Future Research

In this study, we analyze GHC fellowship selection process to investigate the roles of applicants' and reviewers' demographics and position characteristics on the evaluation and whether there exist biases. We find that female applicants who previously applied and had worked in public health are reviewed more positively. Reviewers tend to favor applicants with the same citizenship, and have a higher standard when they are familiar with the skills required by the position and when they review for their home country. Reviewers of the same gender and/or citizenship are more likely to select the same candidates as semifinalists.

To reduce bias, we recommend that normalized scores should be used when comparing evaluations from different reviewers and the review team should consist of reviewers of different gender and citizenship, with at least one who is familiar with the required skills. Assigning reviewers by their review preferences can effectively reduce the inconsistency of their evaluations. For Round 3, we find that reviewers select candidates who are recommended by two reviewers and the rest by ranking their average scores. We propose data-driven methods such as normalized score ranking and random forest ensemble to not only improve accuracy and consistency of the final selection, but also to reduce labor costs and processing time. All in all, developing and setting a clear criteria rule beforehand can effectively remove bias from the evaluation process.

Future research should further investigate features that we did not include in our analysis such as age and language proficiency. Matched characteristics can be extended from binary to continuous (e.g., how many skills the two reviewers have in common). Additional date/time information of each observation will allow for a time series analysis (e.g., how reviewers change their standards over time). Whether or not the process is blinded is very important. In other words, knowing the set of information available to the reviewers can enhance our understanding of biases. Lastly, incorporating finer details about applicants' abilities (e.g., standardized test scores or performances in subsequent rounds of interviews) will allow us to better quantify the performance of existing review process and recommend appropriate strategies for improvement.

References

[1] R. Pingitore, B. L. Dugoni, R. S. Tindale, and B. Spring, "Bias against overweight job applicants in a simulated employment interview.," *Journal of Applied Psychology*, vol. 79, no. 6, p. 909, 1994.

[2] S. L. S. Purkiss, P. L. Perrewé, T. L. Gillespie, B. T. Mayes, and G. R. Ferris, "Implicit sources of bias in employment interview judgments and decisions," *Organizational Behavior and Human Decision Processes*, vol. 101, no. 2, pp. 152–167, 2006.

[3] L. Bornmann, R. Mutz, and H.-D. Daniel, "Gender differences in grant peer review: A meta-analysis," *Journal of Informetrics*, vol. 1, no. 3, pp. 226–238, 2007.

[4] R. Smith, "Peer review: a flawed process at the heart of science and journals," *Journal of the royal society of medicine*, vol. 99, no. 4, pp. 178–182, 2006.

[5] M. Teplitskiy, D. Acuna, A. Elamrani-Raoult, K. Kording, and J. Evans, "The social structure of consensus in scientific review," *arXiv preprint arXiv:1802.01270*, 2018.

[6] H. Bozdogan, "Model selection and akaike's information criterion (aic): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[7] P. J. Bickel, E. A. Hammel, and J. W. O'Connell, "Sex bias in graduate admissions: Data from berkeley," *Science*, vol. 187, no. 4175, pp. 398–404, 1975.

A Supplementary Materials

A.1 Details of Data Pre-Processing

Text fields are pre-processed to remove all possible typos such as *Viet Nam* to *Vietnam* and replace abbreviations by full names such as *U.S.* to *United States*. We compute binary features based on matched characteristics, including whether the two reviewers share the same citizenship, whether the reviewer and the applicant share the same citizenship, and whether the two reviewers have any skills in common. Citizenship and skill sets can have multiple values. When we compute binary features, we consider all possible overlapping values. If any of the values match, we will compute as it is overlapped.

Normalized score is calculated for each total score to scale it to the neutral standard. Denote the total score an applicant i receives from a reviewer j as s_i^j. We define normalized score as:

$$s^j_{norm_i} = \frac{s_i^j - \min_k(s_k^j)}{\max_k(s_k^j) - \min_k(s_k^j)},$$

where k represents any applicant reviewed by reviewer j. The distribution of normalized scores is as follows: minimum is 0, maximum is 1, first quartile is 0.3158, median is 0.5385, mean is 0.5272, third quartile is 0.7500, and standard deviation is 0.2739.

We clean the rankings of applicants as follows: semifinalists retain their ranks (generally 1-10), and alternates have an updated rank as the sum of their original rank and the last rank of the semifinalists. Applicants who are not selected as neither semifinalists nor alternates are ranked zero in the data, so we assign their ranks as the number of applicants to reflect the fact that the larger the rank, the less favorable they are.

In Round 3, we only observe the reviewer ID of the candidates who are eventually selected in this round. In other words, we do not have information about who review the applicants that are not successful in Round 3. We assume that Round 3 reviewers review every applicant recommended by Round 2 reviewer(s) for the same position. Then, we can impute the missing Round 3 reviewer ID by matching the position ID and applicant ID from Round 2.

Table 1: Estimates from different regression models of the success metrics of applicants

	Dependent variable:				
	Rank	Normalized Score		Semifinalist	Passed
	negative binomial	*beta*	*panel linear*	*logistic*	*logistic*
	(1)	(2)	(3)	(4)	(5)
Male applicants	0.056**	−0.151***	−0.036***	−0.191***	−0.156***
	(0.025)	(0.030)	(0.007)	(0.057)	(0.052)
Worked in public health	−0.041***	0.117***	0.027***	0.139***	0.120***
	(0.012)	(0.014)	(0.003)	(0.026)	(0.024)
Eligible	−0.108***	−0.116***	−0.105***	0.348***	0.407***
	(0.033)	(0.039)	(0.014)	(0.075)	(0.069)
Previously applied	−0.110***	0.174***	0.043***	0.326***	0.322***
	(0.028)	(0.033)	(0.007)	(0.060)	(0.056)
Race:Asian	−0.063	0.060	0.030***	0.212**	0.201**
	(0.040)	(0.048)	(0.011)	(0.088)	(0.082)
Race:Hispanic	−0.043	0.004	0.022	0.217	0.140
	(0.064)	(0.075)	(0.018)	(0.138)	(0.130)
Race:White	−0.039	−0.042	0.005	0.102	0.135**
	(0.031)	(0.037)	(0.009)	(0.068)	(0.063)
Same citizenship	−0.145***	0.041	0.019	0.416***	0.512***
	(0.032)	(0.038)	(0.017)	(0.072)	(0.066)
Skilled reviewer	0.092***	0.012	0.022	−0.240***	−0.322***
	(0.023)	(0.027)	(0.022)	(0.049)	(0.046)
Review for placement	0.043	−0.003	0.041**	−0.106*	−0.157***
	(0.027)	(0.032)	(0.020)	(0.058)	(0.055)
Review for home	0.099***	0.059**	0.074***	−0.332***	−0.325***
	(0.024)	(0.028)	(0.026)	(0.052)	(0.048)
Male reviewer	−0.066**	−0.072**		0.177***	0.195***
	(0.026)	(0.030)		(0.056)	(0.052)
Constant	4.415***	0.044		−1.474***	−0.981***
	(0.042)	(0.049)		(0.097)	(0.088)
Observations	8,440	8,440	8,440	8,440	8,440
R^2		0.013	0.8215		
Adjusted R^2			0.8159		
Log Likelihood	−44,359.210	856.385		−4,925.507	−5,517.706

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: Estimates from Logit model to predict round 3 selection when one reviewer suggests in round 2

	Dependent variable:
	Round 3 selected
Average normalized scores	0.172
	(0.775)
Maximum normalized scores	2.644***
	(0.908)
Maximum score	−0.0003
	(0.040)
Same citizenship reviewer 1, 3	−0.203
	(0.215)
Same citizenship reviewer 2, 3	−0.025
	(0.192)
Same citizenship reviewer 3, applicant	0.360
	(0.226)
Constant	−2.495***
	(0.753)
Observations	663
Log Likelihood	−431.385
Akaike Inf. Crit.	876.770
Note:	*p<0.1; **p<0.05; ***p<0.01

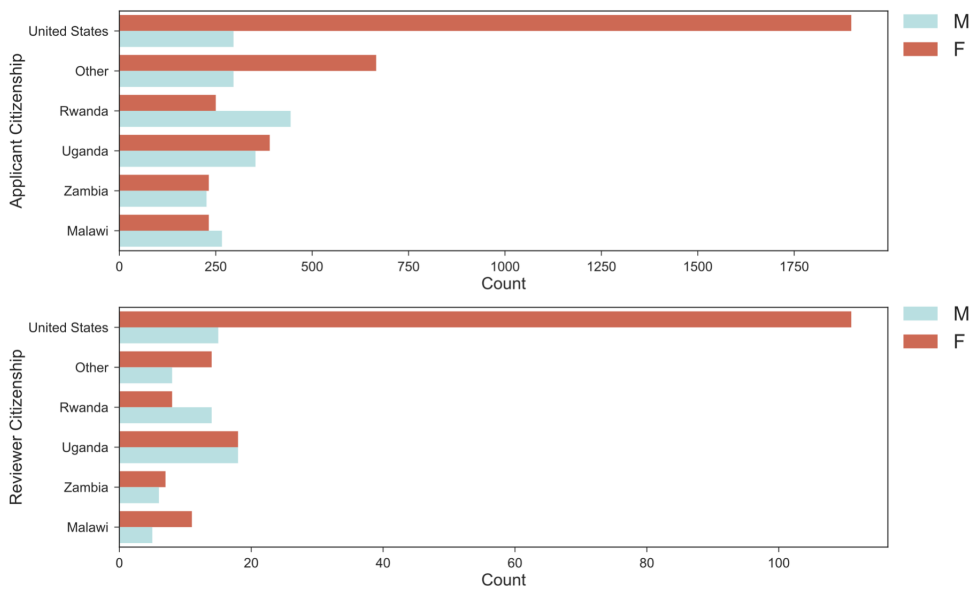

Figure 1: **Distribution of citizenship and gender in applicants and reviewers**. We plot number of applicants and reviewers based on their citizenship and gender. Majority of applicants and reviewers are from United States. Most applicants and reviewers are female.

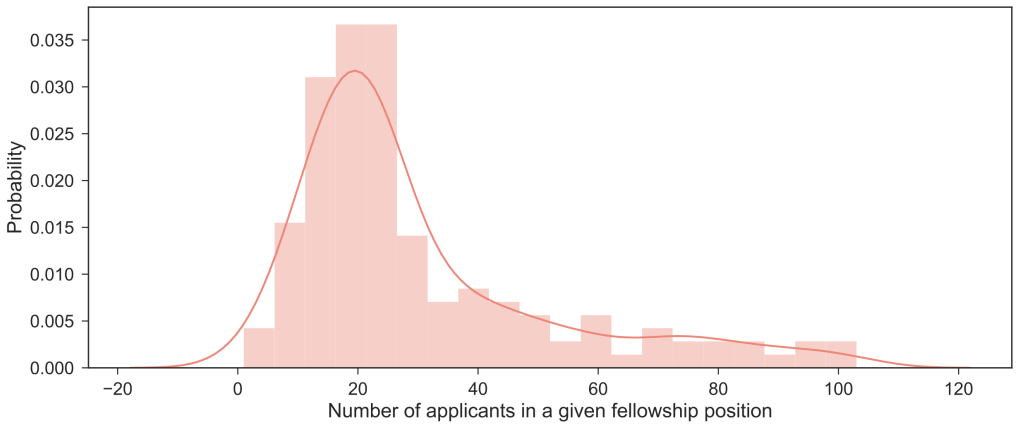

Figure 2: **Distribution of number of applicants for a given position**. This is a normalized histogram of number of applicants applying to GHC fellowship positions.

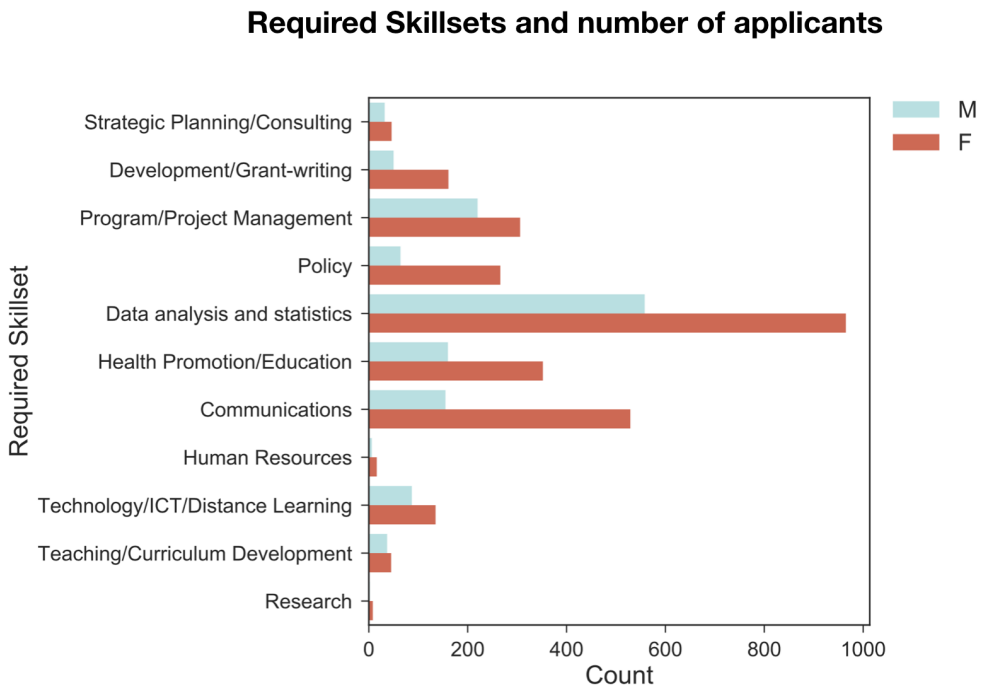

Figure 3: **Required skill sets and number of applicants**. We plot number of applicants applying to jobs classified by required skill sets (Blue for male and red for female). Data analysis and statistics skill is the most required skill whereas research skill is the least required skill set.

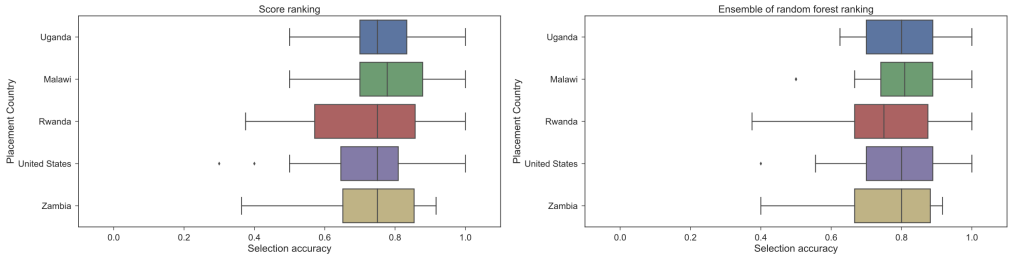

Figure 4: **Round 3 selection accuracy based on placement country**. We show bar plot of selection accuracy produced by proposed metric compared to round 3 selection grouping by placement country. Left panel shows accuracy distribution performed by score ranking model. The ranking is calculated based on consensus of reviewers round 2 follows by maximum normalized score between reviewers. Right panel shows accuracy of random forest-based (RF-based) ensemble to predict selection probability in round 3. The models are trained based on sample 30 percent of the dataset. For each job position, we then rank applicants by average predicted probability from ensemble of random forest models. Accuracy of is calculated by an overlapping between group applicants with highest probability and round 3 selection. We get global accuracy of 73.84% and 77.32% using score ranking and RF-based ensemble respectively.