# Supplementary Material: Studying Bias in GANs through the Lens of Race

Vongani H. Maluleke*, Neerja Thakkar*, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A. Efros, Angjoo Kanazawa, and Devin Guillory

UC Berkeley

In this supplementary material document, we first discuss the selected perceived race label categorization. We then describe in detail the three Amazon Mechanical Turk tasks and implementation information. Followed by an analysis of the performance of both AMT annotations and the automatic classifier in evaluating perceived race. Visualizations of various truncation levels are shown next, and finally, we present more details and visualizations of quality ranking.

## 1 Race Labels Categorization

We start with the the FairFace dataset labels, and then collect annotations based on our own condensed categorization. The FairFace dataset started with the commonly accepted race categories from the U.S. Census Bureau—white, Black, Asian, Hawaiian and Pacific Islanders (HPI), Native Americans (NA), and Latino. They dropped the HPI and NA categories due to insufficient image examples, and expanded the Asian category into four distinct subgroups: Middle Eastern, East Asian, Southeast Asian, and Indian [2]. To reduce perceptual ambiguity (see main paper in section 4.1), we condense the race class labels from seven FairFace classes to three classes—Black, white, and Non-Black or Non-white—where Non-Black or Non-white comprises the Middle Eastern, East Asian, Southeast Asian, Latino Hispanic, and Indian as labeled by FairFace. We also relabel all the images we analyze using our own annotation protocol with three categories and a "Cannot Determine" category.

## 2 Amazon Mechanical Turk (AMT) Details

Amazon Mechanical Turk (AMT) was used to collect annotations for three label tasks, namely; (1) race classification, (2) quality classification, and (3) quality ranking. As mentioned in the paper, these tasks consist of the following questions:

1. **Race Classification:** What is the race of the person in the image?
2. **Real/Fake Classification:** Is this image real or fake?
3. **Image Quality Ranking:** Which image is more likely to be a fake image?

*Implementation details.* Our label tasks were deployed using a custom framework for deploying AMT tasks using our dynamically populated HTML/JavaScript template and the Python API Boto3[1]. Our code enables creating human intelligence tasks (HITs) that show images with a corresponding question, and then the annotator completes a forced-choice answer among a set of specified choices. For quality control, we use both **accuracy** and **consistency** checks. As an accuracy test, workers must get eight of these hidden questions correct. As a consistency test, we duplicate these ten test cases and scatter them throughout the HIT, and the worker must be consistent for eight of the repeated examples. Our hidden test cases are chosen to be adequately obvious such that diligent workers will successfully pass them. If less than eight are answered correctly, the worker's responses are discarded. Furthermore, we ensure that three unique workers answer each question. Each HIT starts with a consent form and a comprehensive description of the task with practice examples with accompanying answers and descriptions; this helps ensure annotators understand the tasks so they can pass our quality control checks.

Next, we go through each of the three tasks and provide qualitative examples for the questions being asked. In total, we asked a total of 50K questions in 1422 unique HITs, where each was labeled by 3 different workers.We did not collect the demographics or other information on AMT workers. Overall, 59 annotators participated in our tasks, and we paid on average $8.71 per hour. Note that *we will release our code used to conduct our experiments for the benefit of the computer vision community.*

### 2.1 Race Classification

The Race Classification AMT task asked the workers to identify the race of the person(s) in the image, by selecting one of the options described below:

1. Black - This is an image of a Black person.
2. white - This is an image of a white person.
3. Non-Black or Non-white - I know the race of the person and the person is not Black or white.
4. Cannot Determine - I cannot tell the race of the person.

The AMT workers were given examples of images and corresponding race class labels- 1, as well as a demo of the interface before they could start the task. Fig. 2 is an example of the deployed Race Classification AMT interface.

### 2.2 Quality Classification

The Quality Classification AMT task asked workers to identify real photographs and fake images by selecting one of the options described below:

1. Real Photograph - This is a photograph of a real person taken using a camera.

[1] `https://boto3.amazonaws.com/v1/documentation/api/latest/index.html`

(a) Black (b) White

(c) Non-Black or Non-white (d) Cannot Determine

Fig. 1: Sample of the race classification examples with corresponding race class labels given to AMT workers before they started Task 1.

2. Fake/Manipulated Image - This is a computer-generated image of a person(s) who do not exist.

To further assist the workers in understanding the difference between the two options, the definition of each of the two options was provided to workers as "Real photographs are images of real person(s) captured using a camera" and "Fake/Manipulated images are computer generated images of a person(s) who do not exist". Fig. 3 is an example of the Quality Classification AMT interface.

**Results from Real/Fake Classification** Using the Real/Fake indicator as a proxy for image quality, we are unable to determine any significant distinctions in generated image quality with respect to race. We use the label "fake" as a proxy for low quality and "real" to represent high quality. This image quality proxy measured was collected by using AMT annotators to determine if an image

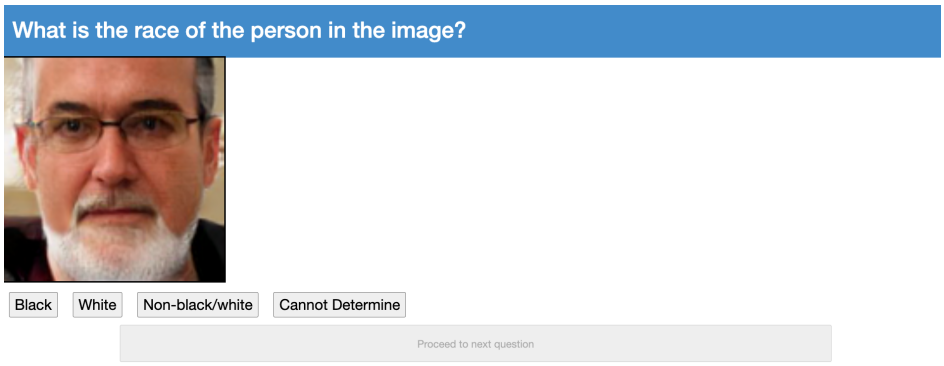



Fig. 2: Race Classification AMT Interface Example.

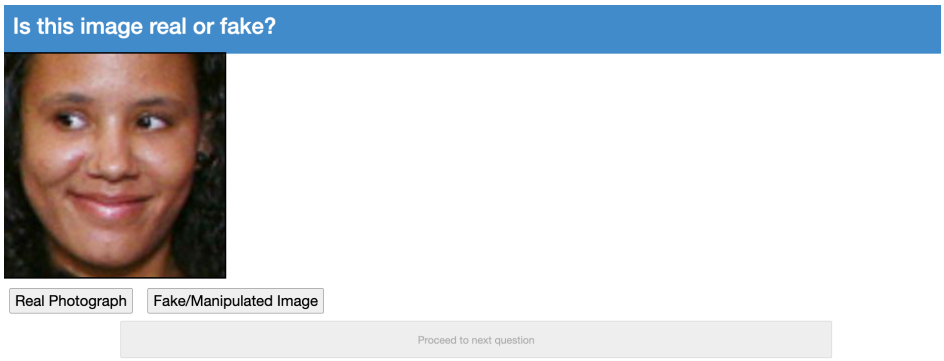



Fig. 3: Quality Classification AMT Interface Example.

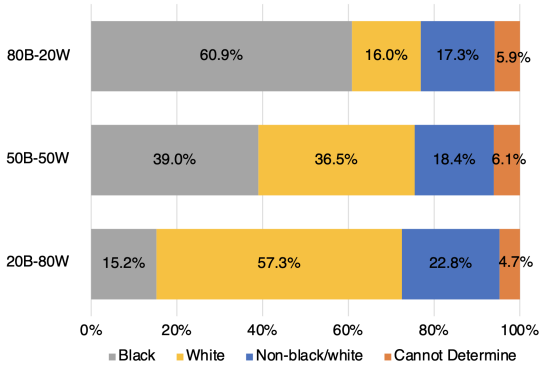
was real or fake. Fig. 4 shows the racial distribution of the generated images that were classified as real/high quality for the three data splits.

In Fig. 4 we observe that the race class ratio of the training images has the same race class ratio as the training data. The race ratio of the 20B-80W, 50B-50W, and 80B-20W training data, respectively, has a race ratio of 0.25, 1, and 4, and the corresponding generated data race ratios of the high quality labeled images are 0.26, 1.1, and 3.8.

### 2.3 Quality Ranking

The Quality Ranking AMT task asked the workers to identify the Fake image between two images by selecting one of the options described below:

1. Image A - Image A is more likely to be a Fake image.




Race Ratios of the Training data and the labeled Generated Images as Real

| Data Split | Training Data | Generated Data |
|---|---|---|
| 80B-20W | 4.0 | 3.8 |
| 50B-50W | 1.0 | 1.1 |
| 20B-80W | 0.3 | 0.3 |

Fig. 4: Distribution of the Generated Images that were labeled as "real", or high quality, images by annotators in the three FairFace data splits (20B-80W, 50B-50W and 80B-20W) with a corresponding table showing the race ratio (Black/white) of the different data splits. This shows that the race ratio of the training data is relatively the same as the race ratio of the generated images that were labeled Real.

2. Image B - Image B is more likely to be a Fake image.

To further assist the workers in understanding what Fake images are, the same definitions from above were given to the workers at the start of the task. Fig. 5 is an example of the Quality Ranking AMT interface.

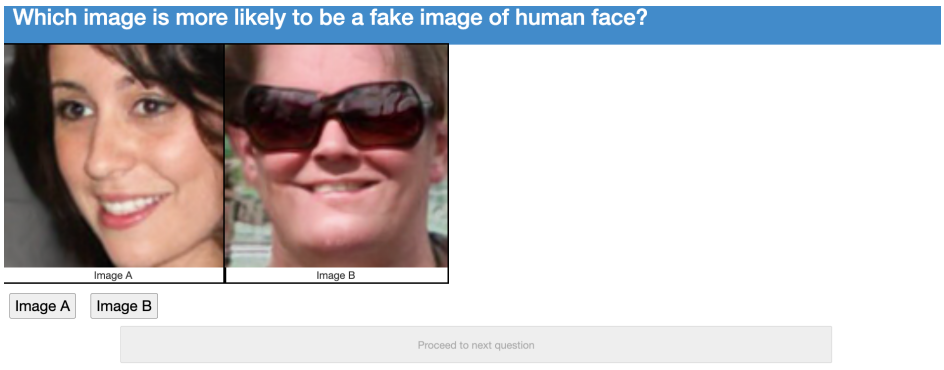



Fig. 5: Quality Ranking AMT Interface Example.

## 3 Race Classification Performance Analysis

In this section we expand on the performance of the AMT annotation compared to the automatic perceived race classifier.

Our perceived race classifier obtained an accuracy of 84%, treating our FairFace labels as ground truth, on the same 1000 images used for human annotation. This gave us more confidence on the performance and validity of the race classifier and its role as a proxy for AMT annotation when conducting our experiments.

The classifier performs better on Black and Other class labels compared to the white class label. The classifier tends to classify white faces as Other which is also slightly observed in the AMT annotations.

Comparing the performance of the human annotation and the perceived race classifier we see that they are both aligned in terms of classifying the different race class labels, and therefore the automatic classifier can be used as a proxy for human annotation. Overall, the classifier outperforms the human annotations. We hypothesize that this could be due to subjective bias present in human annotation, or to the subjective nature of perceived race classification.

## 4 Experimental Results

### 4.1 Relationship between Training and Generated Data Distributions

In this section, we expand on the results that demonstrate that StyleGAN2-ADA's generated data distribution preserves the training data distribution. In the paper, we excluded the "Non-Black or Non-white" and "Cannot Determine" class labels in the generated data to explicitly showcase the ratio of Black and white race class labels in the training and generated data. Fig. 6 (left) shows the generated data distribution with all the class labels where the actual number of the classes are in bracket in the pie charts and Fig. 6 (right) shows distribution for when the "Non-Black or Non-white" and "Cannot Determine" class labels were excluded. To get the generated data distribution with "Non-Black or Non-white" and "Cannot Determine" class labels, these two class labels were dropped and the distribution was recalculated with only white and Black class labels.

### 4.2 Truncation

In order to evaluate properties of truncation, images were generated from StyleGAN2-ADA trained on FFHQ and the 80B-20W, 50B-50W and 80B-20W FairFace-trained generators at truncation levels ranging from $\gamma = 0$ to 1 at intervals of 0.1. Random samples from FFHQ with various levels of truncation can be seen in Fig. 7. As the level of truncation increases, the ratio of faces of people of color to white faces decreases.

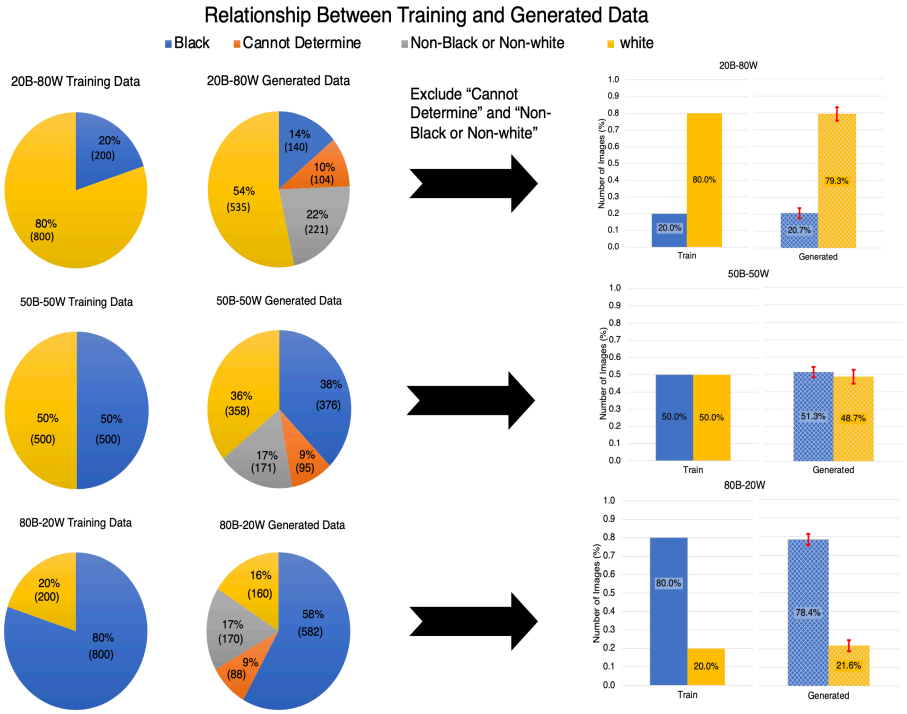



Fig. 6: Racial distribution of training and generated data. Distributions for 20B-80W, 50B-50W and 80B-20W data splits for (left) all the class labels and (right) Black and white class labels only. This figure shows that all of the generative models preserve distribution of the training data.The red bars represents the 95% Wald confidence interval (CI) of the generated data. See Table 1 for the corresponding CI.

### 4.3 Quality Ranking

The raw intra-split pairwise comparison results from our quality ranking experiments can be seen in Tables 2, 3 and 4. From these pairwise comparisons and the inter-split comparisons, we use the `choix` package [3] to produce a Bradley-Terry model that ranks the 3000 images in order of highest quality to lowest perceived image quality. From this global ranking, we visualize the top 25 and bottom 25 images, and also random images from each quartile, which can be seen in Fig. 8.

The raw counts for our top $K$ image compositions per race class label can be seen in Table 6. In order to obtain weighted percentage scores seen in the main paper, the raw counts for the top $K$ images of a particular race class label were first weighted by the expected frequency of the images in each split. This was done by multiplying the raw count by $\frac{1}{2}$, $\frac{1}{5}$, and $\frac{1}{8}$ if the race class label comprised 20%, 50% or 80% of the dataset, respectively. Then, the weighted numbers were divided by the sum of all scores for that particular race class label and given value of $K$.

Table 1: Wald's 95% confidence interval (CI) of generated data from generators trained on FFHQ and 80B-20W, 50B-50W and 80B-20W FairFace .

| Generated Data | Black CI @95% | white CI @95% |
|---|---|---|
| FFHQ | 5.73 ± 1.7 | 94.34 ± 1.7 |
| 20B-80W | 20.70 ± 3.1 | 79.3 ± 3.1 |
| 50B-50W | 51.3 ± 3.7 | 48.7 ± 3.7 |
| 80B-20W | 78.4 ± 3.1 | 21.6 ± 3.1 |

Table 2: **Intra-split pairwise perceived image quality comparison 20B-80W Dataset.**

| More Preferred | Less Preferred | Count |
|---|---|---|
| white | white | 2367 |
| white | Black | 708 |
| white | Other | 982 |
| white | CD | 1129 |
| Black | Black | 186 |
| Black | white | 465 |
| Black | Other | 205 |
| Black | CD | 270 |
| Other | Other | 411 |
| Other | white | 824 |
| Other | Black | 269 |
| Other | CD | 509 |
| CD | CD | 225 |
| CD | Black | 87 |
| CD | white | 230 |
| CD | Other | 133 |

A Bradley-Terry model [1] predicts the probability that a pairwise comparison $i > j$ is true. A ranking of all items can be derived by modeling the probability for pairs in a dataset. The `choix` package [3] produces a Bradley-Terry model by using the Iterative Luce Spectral Ranking algorithm [4]. This algorithm performs maximum-likelihood inference to rank items from a dataset of pairwise comparisons.

Table 3: **Intra-split pairwise perceived image quality comparison 50B-50W Dataset.**

| More Preferred | Less Preferred | Count |
|---|---|---|
| white | white | 990 |
| white | Black | 1305 |
| white | Other | 497 |
| white | CD | 756 |
| Black | Black | 1167 |
| Black | white | 888 |
| Black | Other | 448 |
| Black | CD | 578 |
| Other | Other | 252 |
| Other | white | 475 |
| Other | Black | 599 |
| Other | CD | 323 |
| CD | CD | 159 |
| CD | Black | 274 |
| CD | white | 201 |
| CD | Other | 88 |

Table 4: **Intra-split pairwise perceived image quality comparison of 80B-20W Dataset.**

| More Preferred | Less Preferred | Count |
|---|---|---|
| white | white | 195 |
| white | Black | 798 |
| white | Other | 216 |
| white | CD | 319 |
| Black | Black | 2595 |
| Black | white | 660 |
| Black | Other | 705 |
| Black | CD | 1064 |
| Other | Other | 216 |
| Other | white | 216 |
| Other | Black | 918 |
| Other | CD | 295 |
| CD | CD | 219 |
| CD | Black | 421 |
| CD | white | 83 |
| CD | Other | 80 |

Fig. 7: **FFHQ samples with various levels of truncation.** (top left) truncation of $\gamma = 1$ (no truncation), (top right) truncation of $\gamma = 0.8$, (middle left) truncation of $\gamma = 0.6$, (middle right) $\gamma = 0.4$, (bottom left) truncation of $\gamma = 0.2$, and (bottom right) $\gamma = 0$ (full truncation). As the amount of truncation increases, racial diversity decreases, resulting in an increasingly larger proportion of white faces.

Table 5: **Race label breakdown of global ranking.** For each quartile of the global ranking of 3000 FairFace generated images compiled from different data splits, the percentage of faces annotated as white, Black, non Black/white, and Cannot Determine. white faces are over-represented in the top half of the quality ranking, and under-represented in the bottom half.

| Quartile | white % | Black % | Non Black/white % | Cannot Determine % |
|---|---|---|---|---|
| Top | 48.4 | 26.3 | 22.5 | 2.8 |
| Second | 40.0 | 37.7 | 19.3 | 4.0 |
| Third | 29.7 | 41.3 | 17.2 | 11.7 |
| Bottom | 15.2 | 32.1 | 11.6 | 41.1 |

Table 6: **Top $K$ image composition per-race, raw counts.** Given a ranking of Black and white images across all data splits, we break down the data split that each image came from. The highest quality images ($K = 10, 25, 50$) are more likely to come from a data split where they are over-represented or represented in parity. These are the raw numbers, before normalization.

| **white** | | | |
|---|---|---|---|
| $K$ | 80B-20W | 50B-50W | 20B-80W |
| 10 | 0 | 2 | 8 |
| 25 | 0 | 7 | 18 |
| 50 | 4 | 16 | 30 |
| 100 | 10 | 39 | 51 |
| 500 | 68 | 181 | 251 |

| **Black** | | | |
|---|---|---|---|
| $K$ | 80B-20W | 50B-50W | 20B-80W |
| 10 | 7 | 2 | 1 |
| 25 | 19 | 4 | 2 |
| 50 | 31 | 16 | 3 |
| 100 | 57 | 31 | 12 |
| 500 | 275 | 161 | 64 |

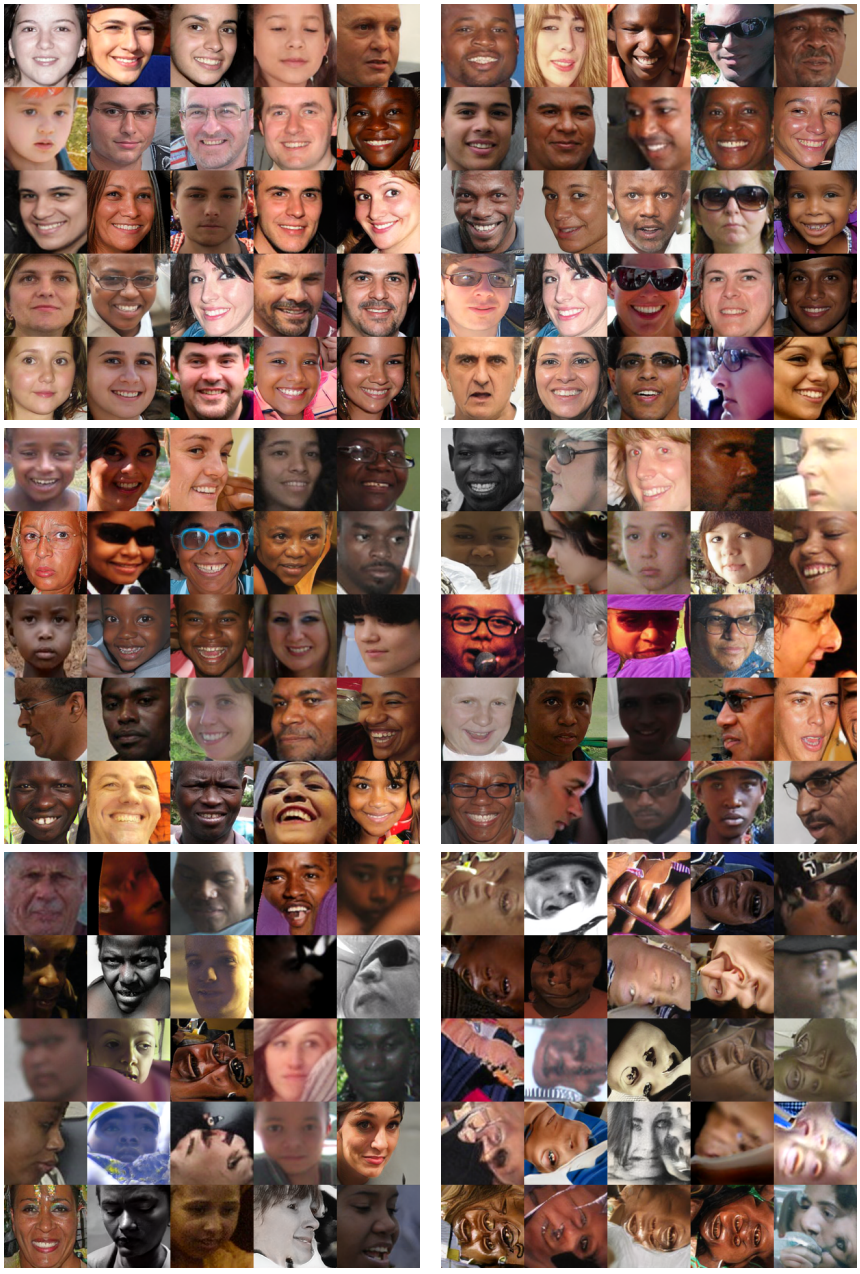


Fig. 8: **Results of global quality ranking across the three FairFace data splits.** Each row represents a particular image quality ranking: (top left) top 25 images in the quality ranking, (top right) samples from the top quartile of images, (middle left) samples from the second top quartile of images, (middle right) third quartile of ranked images, (bottom left) bottom quartile of ranked images, and (bottom right) the bottom 25 images in the quality ranking. As shown in Table 5, white faces are over-represented in the top half of the quality ranking and under-represented in the bottom half.

## References

1. Caron, F., Doucet, A.: Efficient bayesian inference for generalized bradley–terry models. Journal of Computational and Graphical Statistics **21**(1), 174–196 (2012)
2. Kärkkäinen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1547–1557 (2021). https://doi.org/10.1109/WACV48630.2021.00159
3. Lucas, M., Brendan, H.: Github (2022), `https://github.com/lucasmaystre/choix.git`
4. Maystre, L., Grossglauser, M.: Fast and accurate inference of plackett–luce models. Advances in neural information processing systems **28** (2015)