Appendix to Fair Pairs: Fairness-Aware Ranking Recovery from Pairwise Comparisons

A Expected Probabilities under the Bradley-Terry-Luce model

Proposition 1 (Stronger Individuals Winning)**.** *For two individuals i and j with average perceived scores s_i and s_j, the expected probability of the stronger individual winning a pairwise comparison is:*

$$\mathbb{E}(p_{\text{stronger}}) = \iint \frac{e^{\max(s_i,s_j)}}{e^{s_i}+e^{s_j}}\, p(s_i)p(s_j)\, ds_j ds_i \tag{1}$$

where $p(s_i)$ and $p(s_j)$ are informed by the perceived score distributions.

Let s be a random variable for the average perceived score, with probability density function $p(s)$. Let us randomly pick two individuals i, j independently. The "stronger" of the two individuals has the average perceived score $\max(s_i, s_j)$. Then, the probability of the stronger of the two individuals winning a pairwise comparison under the Bradley-Terry-Luce model [1] is a new random variable given by:

$$p_{\text{stronger}} = \frac{e^{\max(s_i,s_j)}}{e^{s_i}+e^{s_j}} \tag{2}$$

The expected value of a random variable Y depending on random variables X_1, X_2 is given by:

$$\mathbb{E}(Y) = \iint y(x_1,x_2) f(x_1,x_2)\, dx_1 dx_2 \tag{3}$$

where $f(x_1, x_2)$ is the joint probability density function. Since the two individuals are picked independently, the joint probability density function of their average perceived scores is given by $p(s_i)p(s_j)$. Therefore, the expected probability of a stronger individual winning is given by:

$$\mathbb{E}(p_{\text{stronger}}) = \iint \frac{e^{\max(s_i,s_j)}}{e^{s_i}+e^{s_j}}\, p(s_i)p(s_j)\, ds_j ds_i \tag{4}$$

In our setup, the average perceived scores of the privileged group are equal to its skill scores, and the average perceived scores of the unprivileged group are equal to its skill score plus bias. Both skill scores and bias are assumed to be normally distributed and independent. Thus, the average perceived score is normally distributed as well, with its mean being the sum of the means of the skill score and bias distributions, and its variance being the sum of the variances accordingly.

As a result, the probability density functions of the average perceived scores differ between the privileged and the unprivileged group. To accommodate for this difference, we calculate $\mathbb{E}(p_{\text{stronger}})$ for three cases: (i) both individuals are from the privileged group, (ii) both individuals are from the unprivileged group, and (iii) one individual is from the privileged and one individual is from the unprivileged group. In our setup, there are 200 individuals in each group, resulting in $200*200 = 40,000$ possible pairs for cases (i) & (ii) and $200 * 200 * 2 = 80,000$ possible pairs for case (iii). We obtain the overall expected probability of a stronger individual winning as the mean of the cases' expected probabilities, weighted by $\frac{1}{4}$, $\frac{1}{4}$, and $\frac{1}{2}$ accordingly. Finally, optimize the parameters for $\mathcal{N}_{\text{skill}}(\mu_{\text{skill}}, \sigma_{\text{skill}})$ and $\mathcal{N}_{\text{bias}}(\mu_{\text{bias}}, \sigma_{\text{bias}})$ to obtain a given value for p_{stronger}.

Proposition 2 (Privileged Individuals Winning)**.** *For any privileged individual with average perceived scores s_p and any unprivileged individual with average perceived scores s_u, the expected probability of the privileged individual winning a pairwise comparison is:*

$$\mathbb{E}(p_{\text{discr}}) = \iint \frac{e^{s_p}}{e^{s_p}+e^{s_u}}\, p(s_p)p(s_u)\, ds_p ds_u \tag{5}$$

where $p(s_p)$ and $p(s_u)$ are informed by the perceived score distributions.

Let s_p be a random variable for the average perceived score of the privileged group, with probability density function $p(s_p)$. Let s_u be a random variable for the average perceived score of the privileged group, with probability density function $p(s_u)$. We can calculate both probability density functions from the parameters given for the normally distributed and independent random variables *skill score* and *bias*. Let us pick one privileged individual p and one unprivileged individual u independently. Then, the probability of the privileged individual winning a pairwise comparison over the unprivileged individual under the Bradley-Terry-Luce model is a new random variable given by:

$$p_{\text{discr}} = \frac{e^{s_p}}{e^{s_p}+e^{s_u}} \tag{6}$$

Following the same argumentation as in the previous derivation, the joint probability function of the individuals' average perceived scores is given by $p(s_p)p(s_u)$, and the expected probability of a privileged individual winning is given by:

$$\mathbb{E}(p_{\text{discr}}) = \iint \frac{e^{s_p}}{e^{s_p}+e^{s_u}}\, p(s_p)p(s_u)\, ds_p ds_u \tag{7}$$

Note that for proposition 2, the probability density functions $p(s_p)$ and $p(s_u)$ are directly defined and we do not need to consider separate cases. We combine the calculations for $\mathbb{E}(p_{\text{stronger}})$ and $\mathbb{E}(p_{\text{discr}})$ into a single optimization problem and determine the parameters for $\mathcal{N}_{\text{skill}}(\mu_{\text{skill}}, \sigma_{\text{skill}})$ and $\mathcal{N}_{\text{bias}}(\mu_{\text{bias}}, \sigma_{\text{bias}})$ accordingly.

B Implementation of the Recovery Algorithms

We used the implementations of David's Score, RankCentrality, and other baseline ranking recovery methods provided in the official GNNRank code repository[1]. We fixed an error in GNNRank's implementation of RankCentrality that would prevent the algorithm from accurately recovering rankings. In addition to these algorithms, we implemented a baseline ranking recovery method that simply assigns random ranks to individuals.

In order to facilitate fast experimentation, we trained GNNRank always only once every graph was strongly connected and then later employed the same model without re-training even after more comparisons were conducted. The original paper highlights such generalization capabilities, and we were able to replicate these findings in preliminary experiments for our synthetically generated data. Only for a few trials did GNNRank not generalize in some iterations. The wrongly recovered rankings result in temporary spikes in the results over time presented in section C and can easily be identified. Since the empirical dataset we evaluated [3] only has a single comparison for each pair of images, GNNRank does not generalize well on the subsampled comparison graph and needed to be re-trained for each evaluation.

Fairness-Aware PageRank does not consider edge weights, i.e. comparisons are either won or lost. We thus had to specify a cutoff for comparisons to consider and comparisons to drop. We set this cutoff to 0.4 for edge weights in $[0, 1]$, allowing for a middle ground of both candidates "winning" if they win approximately the same amount of comparisons. We adopt the LFPR_{P} variant of Fairness-Aware PageRank, as the original authors report good performance on all tests for this variant.

[1] https://github.com/SherylHYX/GNNRank

C Results in Detail

C.1 Simulated Pairwise Comparisons

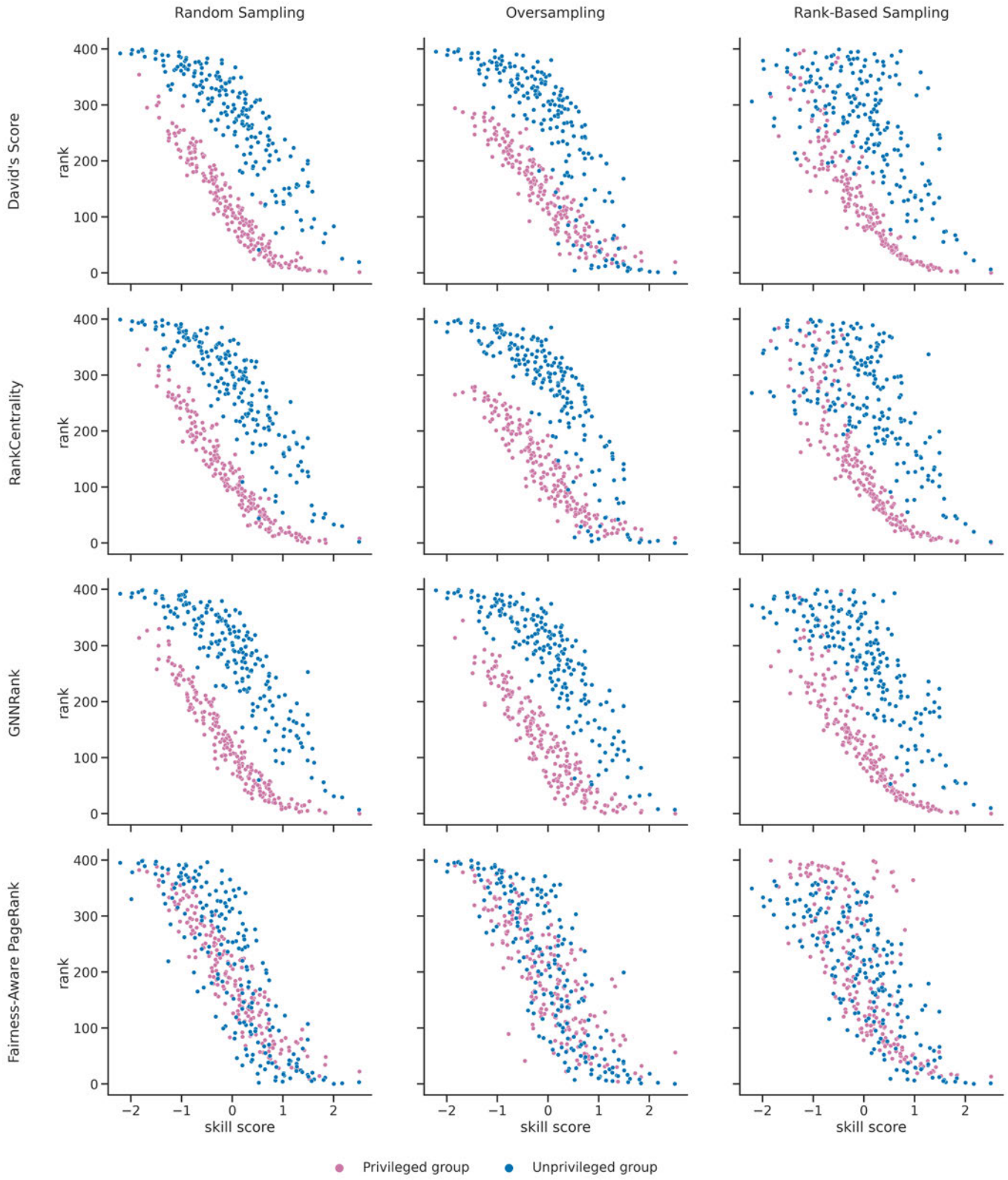

Figure 1. Correlations of Skill Score (higher is better) and Rank (lower is better) by Recovery Method. 400 individuals in 2 equal size groups, after 1000 iterations of pairwise comparisons, using Oversampling. In each iteration 20 %, that is 80 individuals, are compared using the BTL-model [1].

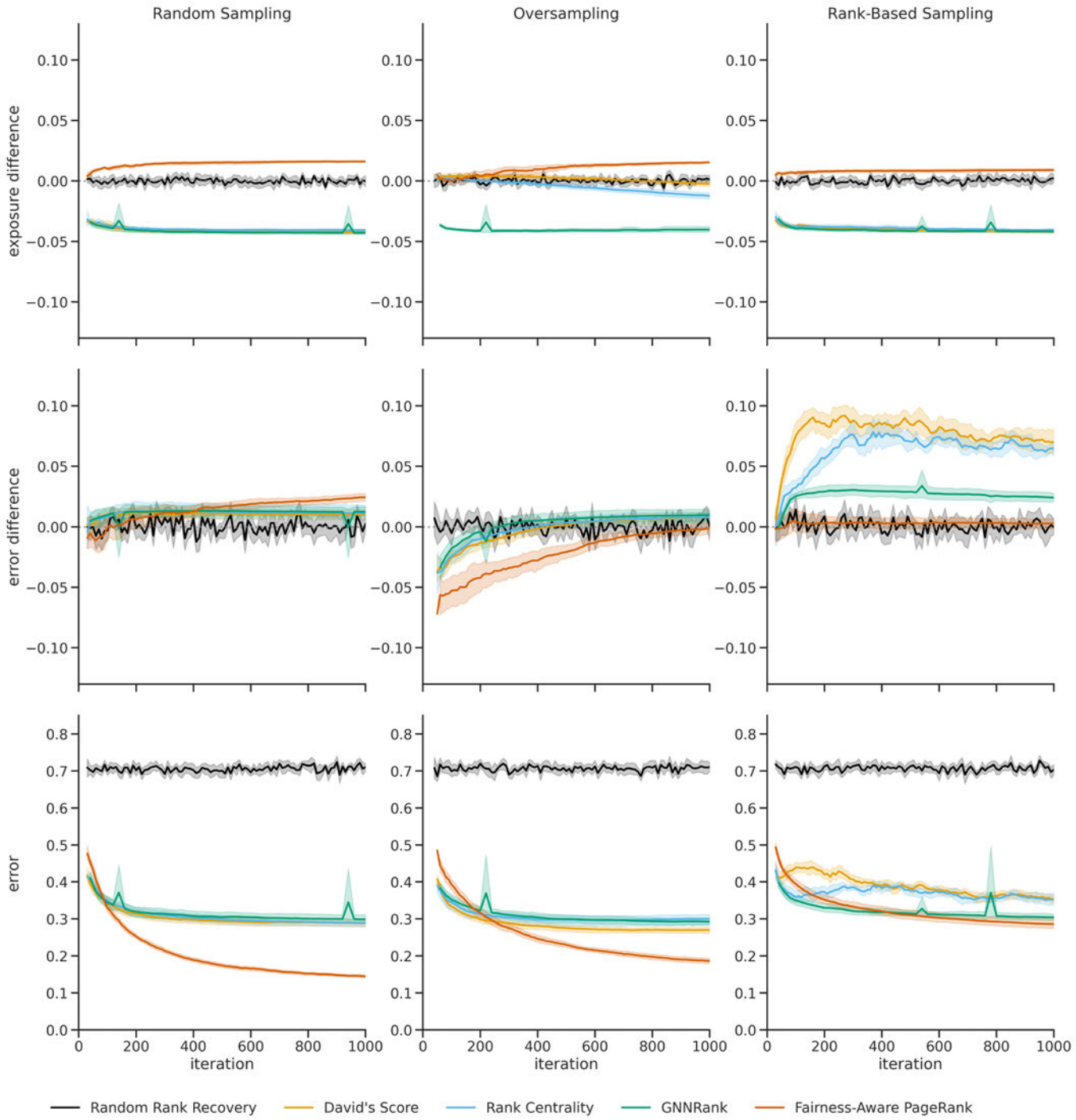

Figure 2. Results from Simulated Pairwise Comparisons, by Sampling Approach (columns) and Ranking Recovery Method (color). Exposure difference (top) and error difference (middle) are group-conditioned measures of fairness, error (bottom) reflects the whole ranking. Iterations of sampling and pairwise comparison on the x-axis.

C.2 *Post-Processing with EPIRA*

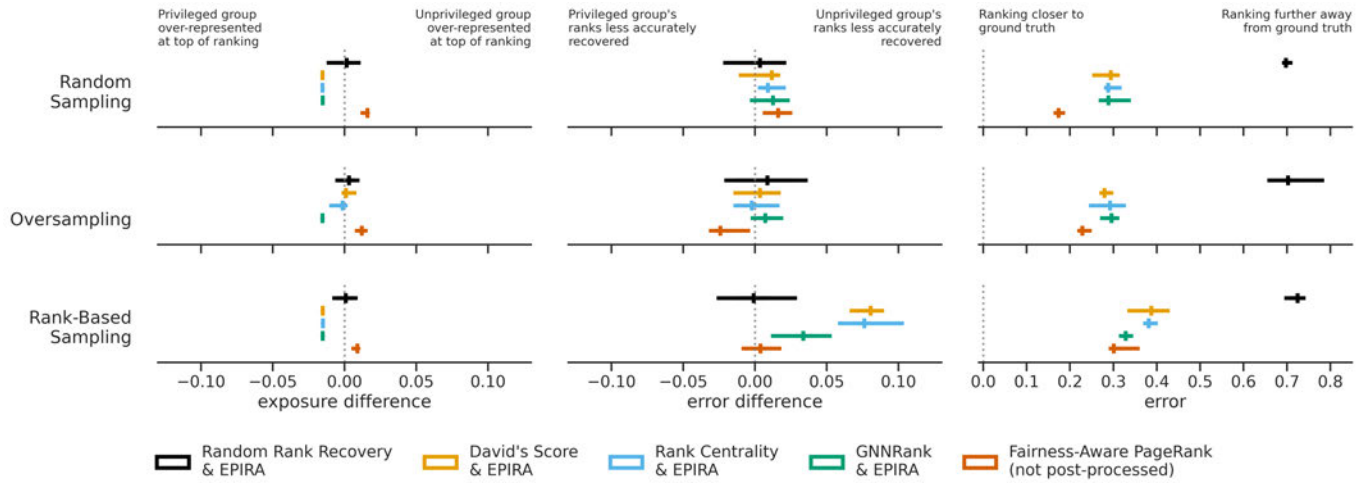

Figure 3. Post-Processed Results after 500 Iterations of Simulated Pairwise Comparisons, by Sampling Approach and Ranking Recovery Method. Post-processing was performed using the EPIRA algorithm [2] with $\mathrm{bnd} = 0.9$, as suggested by the authors. While EPIRA effectively limits exposure difference between the groups, it does not improve error difference under Rank-Based Sampling. Overall error does not see any improvements. Thus, in our simulations, Fairness-Aware PageRank outperforms post-processing with EPIRA.

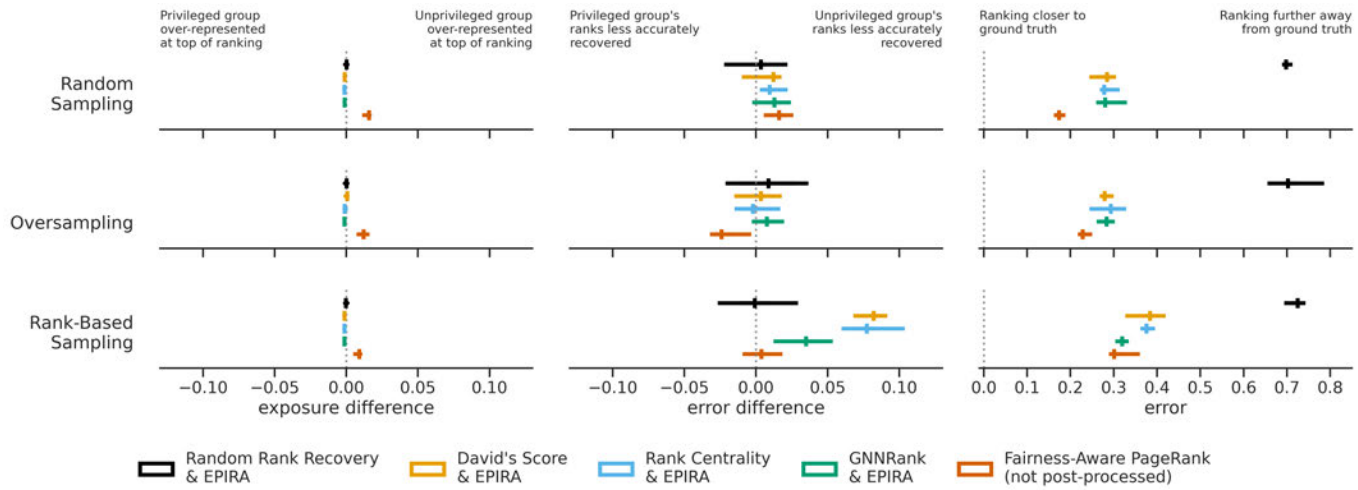

Figure 4. Post-Processed Results after 500 Iterations of Simulated Pairwise Comparisons, by Sampling Approach and Ranking Recovery Method. Post-processing was performed using the EPIRA algorithm [2] with $\mathrm{bnd} = 0.99$. This further improves exposure difference, even beyond the results achieved by Fairness-Aware PageRank. As shown in figure 3, however, error difference and overall error see no improvements.

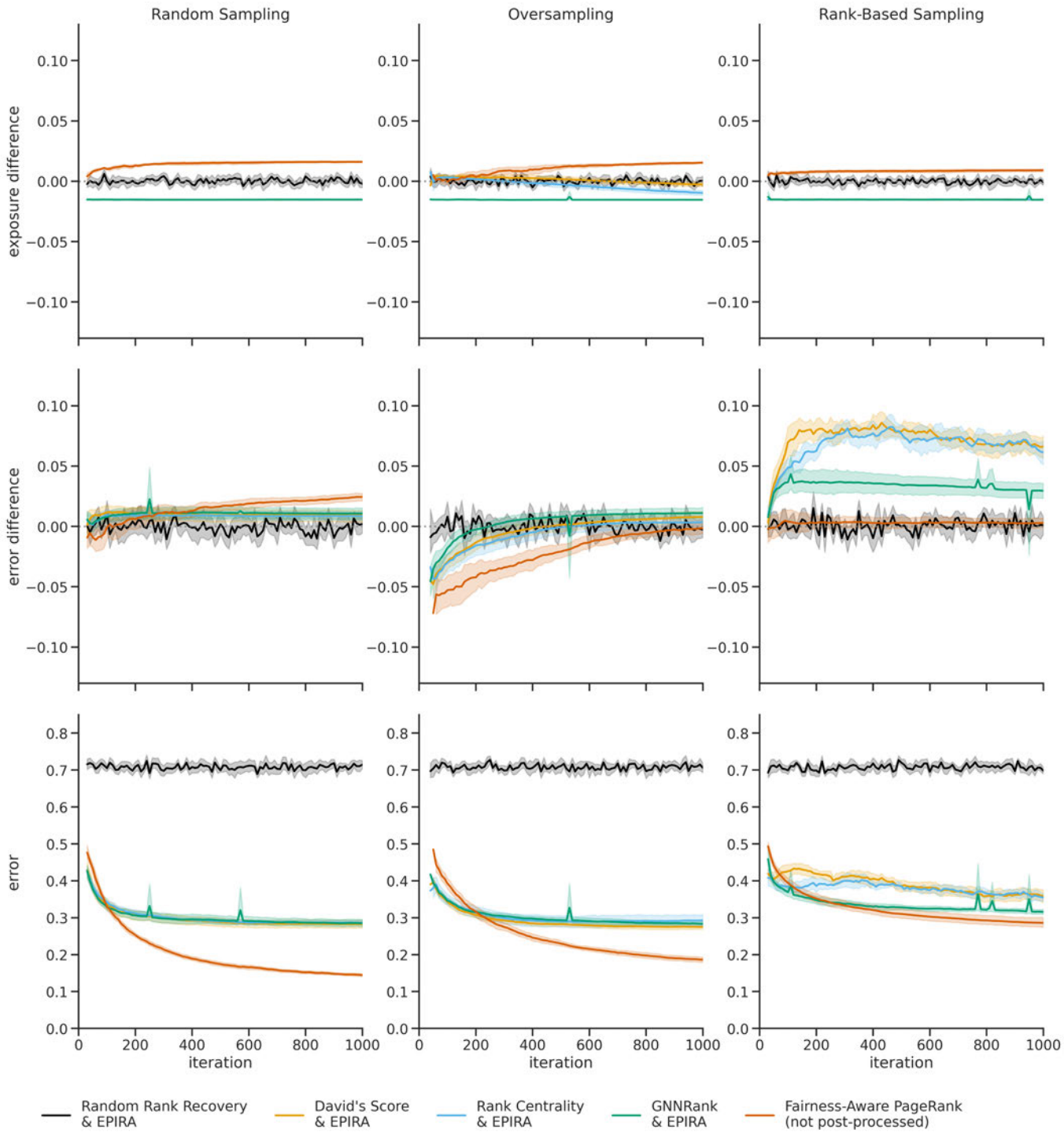

Figure 5. Results from Simulated Pairwise Comparisons, post-processed with EPIRA [2]. Rows indicate measures of fairness (top, middle) and error (bottom). EPIRA was applied with $\mathrm{bnd} = 0.9$, as suggested by the authors.

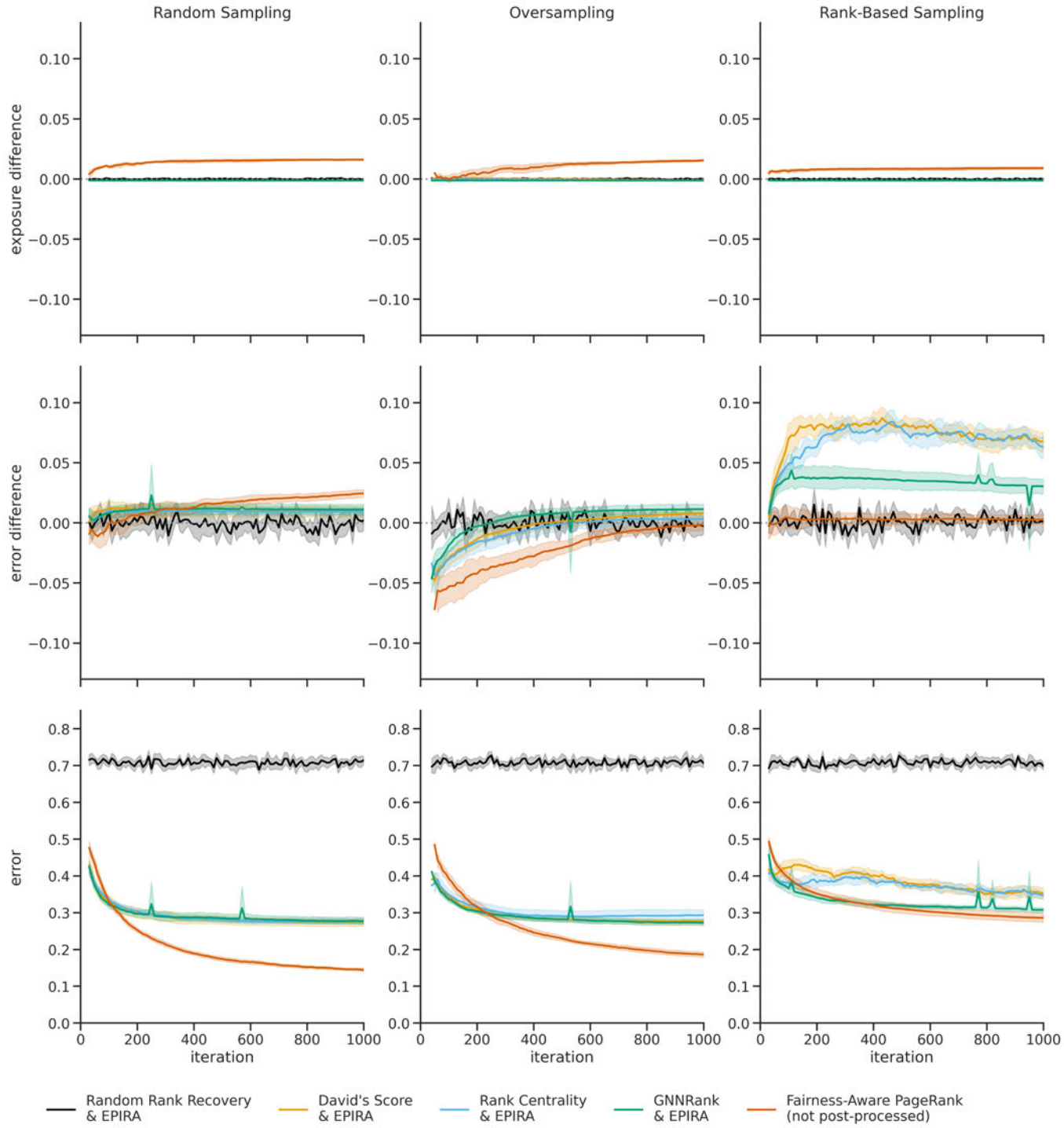

Figure 6. Results from Simulated Pairwise Comparisons, post-processed with EPIRA [2]. Rows indicate measures of fairness (top, middle) and error (bottom). EPIRA was applied with a stricter value for $\mathrm{bnd} = 0.99$.

C.3 *Post-Processing with FA*IR*

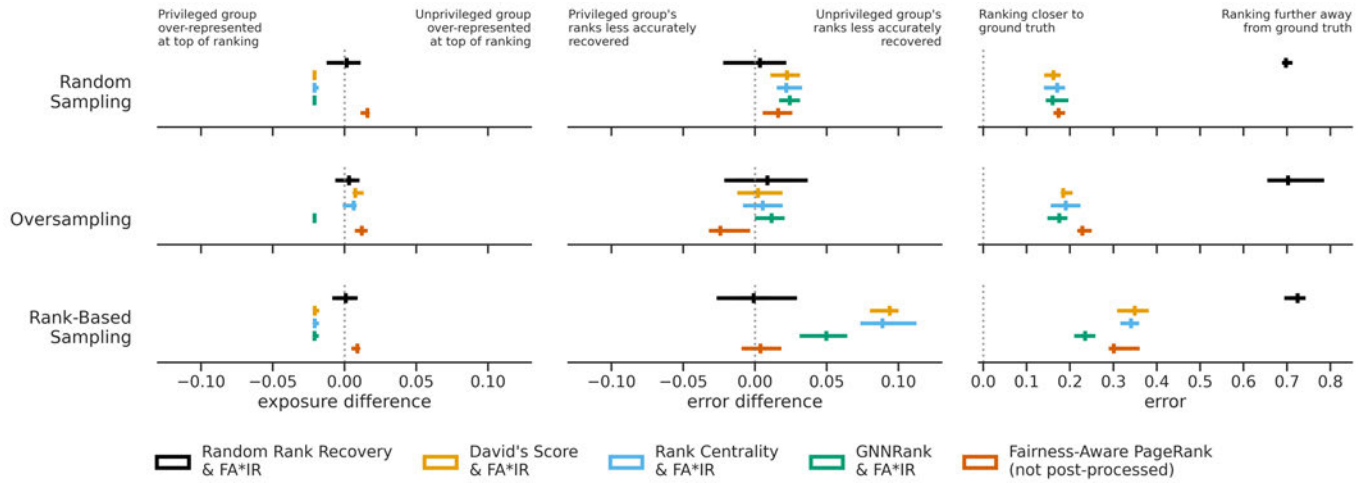

Figure 7. Post-Processed Results after 500 Iterations of Simulated Pairwise Comparisons, by Sampling Approach and Ranking Recovery Method. Post-processing was performed using the FA*IR algorithm [4] with $p = 0.5$ and $\alpha = 0.1$. FA*IR is able to effectively limit exposure difference while no over-shooting the way Fairness-Aware PageRank does. The post-processing technique also improves overall accuracy and is able to outperform Fairness-Aware PageRank, in particular if paired with GNNRank for ranking recovery. FA*IR does, in contrast to Fairness-Aware PageRank, negatively impact error difference.

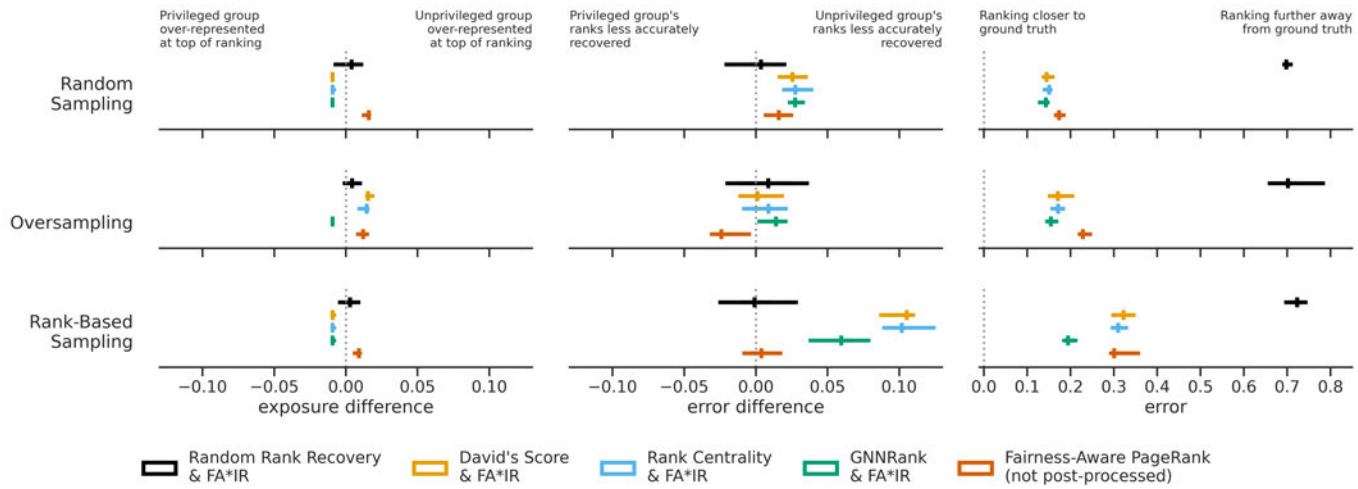

Figure 8. Post-Processed Results after 500 Iterations of Simulated Pairwise Comparisons, by Sampling Approach and Ranking Recovery Method. Post-processing was performed using the FA*IR algorithm [4] with $p = 0.6$ and $\alpha = 0.1$. Increasing the value of p to 0.6 yields additional improvements in exposure difference and overall error, as compared to figure 7.

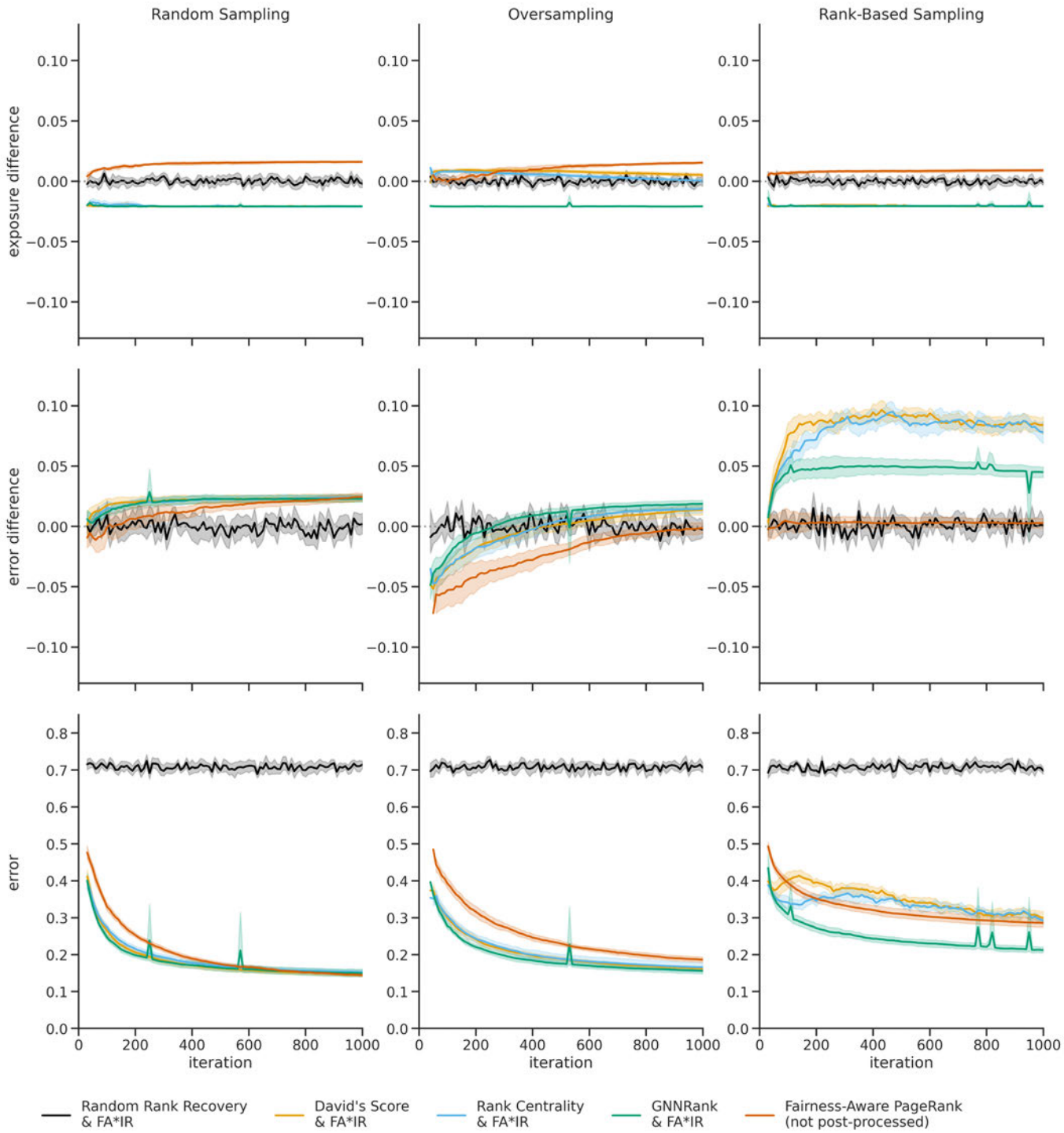

Figure 9. Results from Simulated Pairwise Comparisons, post-processed with FA*IR [4]. Rows indicate measures of fairness (top, middle) and error (bottom). FA*IR was applied with $p = 0.5$ and $\alpha = 0.1$, as suggested by the authors.

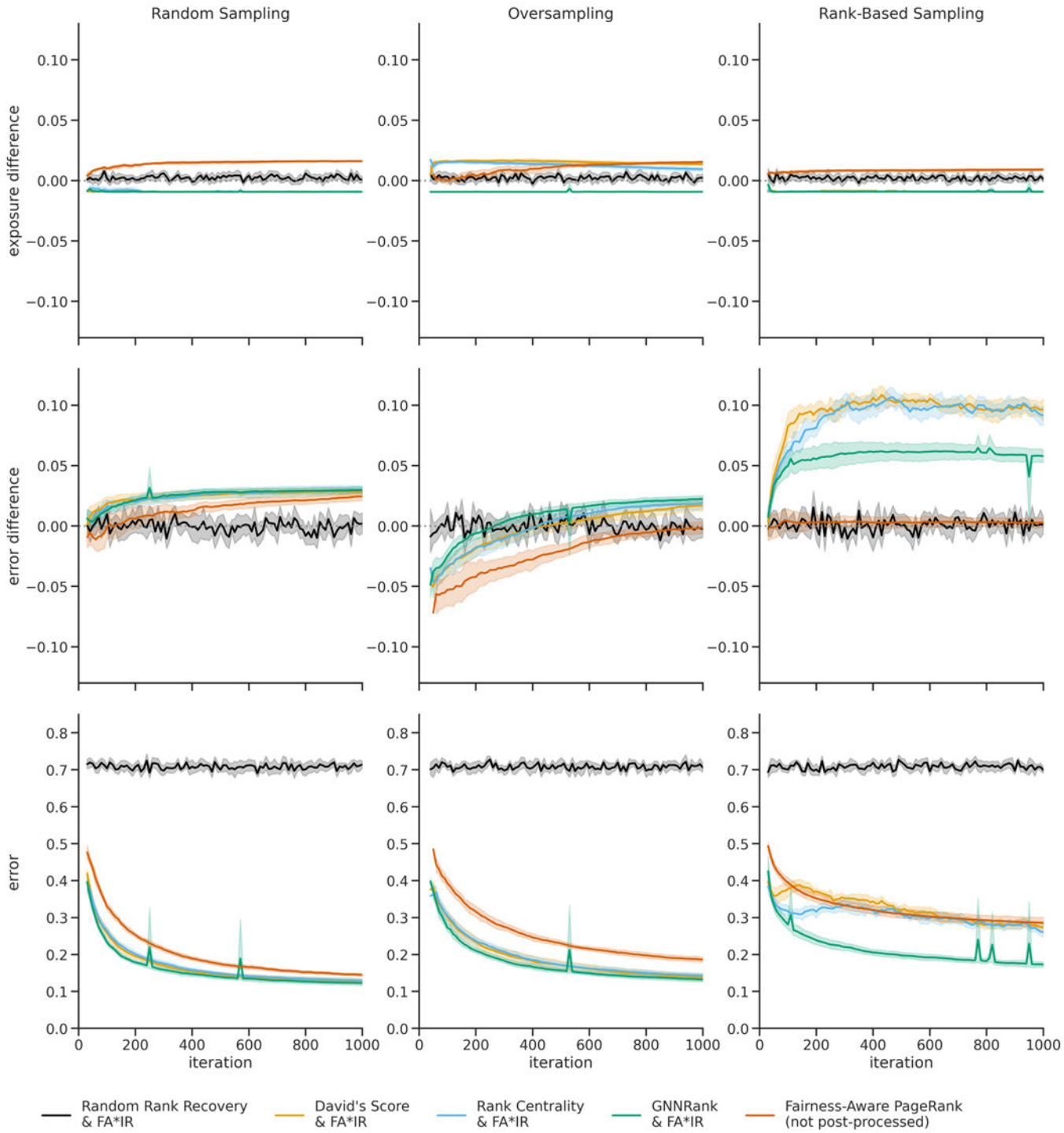

Figure 10. Results from Simulated Pairwise Comparisons, post-processed with FA*IR [4]. Rows indicate measures of fairness (top, middle) and error (bottom). FA*IR was applied with $p = 0.6$ and $\alpha = 0.1$. For our simulated dataset, setting $p = 0.6$ yields both improved exposure difference and error, even if we actually simulate two equal sized groups of individuals.

C.4 *Empirical Data*

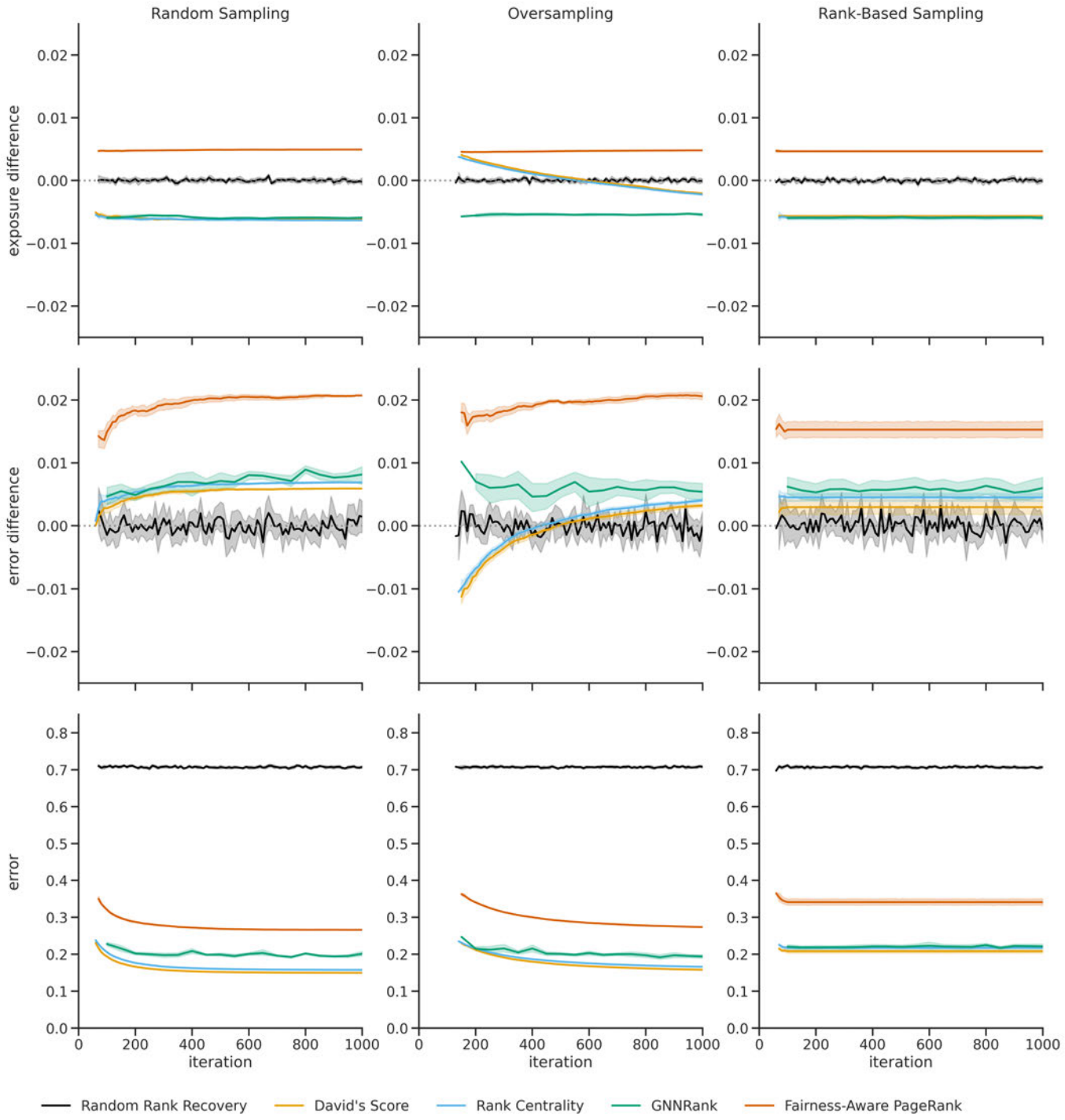

Figure 11. Results from the IMDB-WIKI-SbS dataset [3] with improved labels, by Sub-Sampling Approach and Ranking Recovery Method. Rows indicate measures of fairness (top, middle) and error (bottom). Ranking recovery with GNNRank was only performed every 50 iterations, since on this dataset, GNNRank did not generalize well across iterations and had to be re-trained at each step.

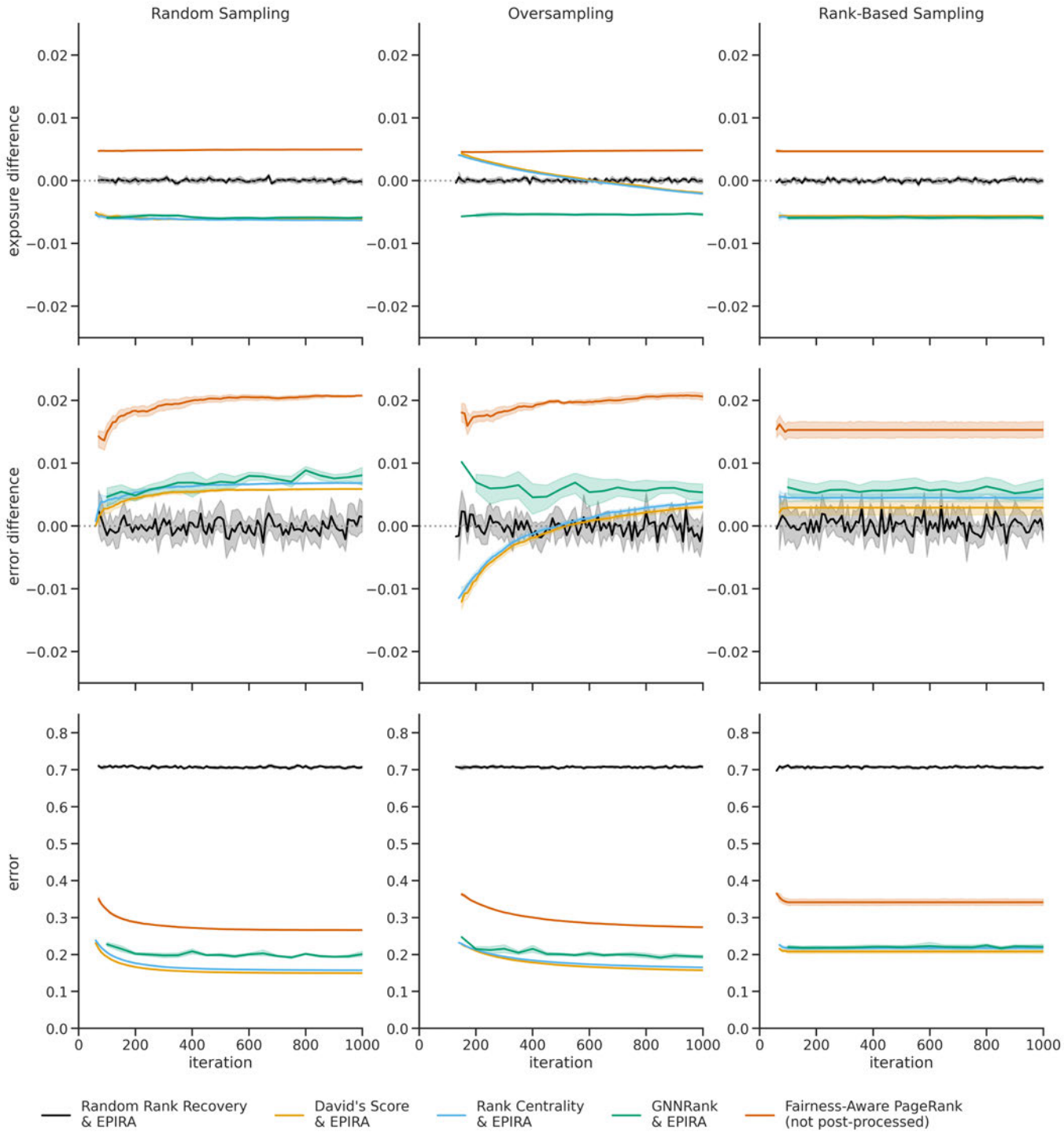

Figure 12. Results from the IMDB-WIKI-SbS dataset [3], post-processed with EPIRA [2]. Rows indicate measures of fairness (top, middle) and error (bottom). EPIRA was applied with a stricter value for $\mathrm{bnd} = 0.99$. Still, it had no effect on the recovered rankings, as exposure differences are considerably smaller in rankings recovered from this emirical dataset as opposed to the synthetic results shown in figure 6.

References

[1] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[2] K. Cachel and E. Rundensteiner. Fairer together: Mitigating disparate exposure in kemeny rank aggregation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1347–1357, 2023.

[3] N. Pavlichenko and D. Ustalov. IMDB-WIKI-SbS: An evaluation dataset for crowdsourced pairwise comparisons. *arXiv preprint arXiv:2110.14990*, 2021.

[4] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. Fa*ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578, 2017.